

Debates over credit for the annotation of genomes

Researchers who devote themselves to sequencing genomes often lack the time to interpret their results. Others don't. The tensions that can result reflect the need for a rethink of sequencers' priorities or a change in approach to collaboration.

It has become a well-established tradition among the publicly funded genome-sequencing community to place sequence data on public and freely accessible databases as the sequences are generated. But there is nothing to stop others from using those data to do good science. Indeed, what else would be the point of a policy of prompt and open access?

But how should credit be attributed? What rights do those doing the sequencing have over subsequent dependent research? And do people putting data and analysis freely on the web prejudice their chances of publication?

To take the last question first: not in the case of *Nature* (or, for that matter, other *Nature* journals). We believe that genomics databases, like preprint servers and conferences, represent a form of intra-community networking from which all researchers benefit. *Nature* does not count them as prior publications. If, exceptionally, an exciting result gets picked up from such a source by the media, that, too, will not necessarily disqualify a paper from consideration, as long as researchers have not pre-empted peer review or publication by encouraging prior publicity. (For a fuller statement of policy, see <http://www.nature.com/nature/author/embargo.html>.) As always, *Nature* will apply its judgement on the significance of a submitted sequence, recognizing that formal peer-reviewed publication of genome maps and sequences represents a necessary culmination of years of research, allowing authors to communicate results and commentary on their significance to the wider world and to gain due credit for their efforts.

This policy applies not only to raw sequence data but also to 'annotations': proposals for the functions of the genes in the database. These often represent substantial pieces of research in themselves. If the results have been added as annotations to a recognized database but have not been subject to the process of peer-reviewed publication, their inclusion on the database is not considered as prior publication by *Nature*.

Publication rights

So far, so good. But the problem, from the point of view of those doing the sequencing, occurs on occasions when they are getting on with their sequencing while others, perhaps better placed to annotate the sequence, are free to use it to publish biologically useful information. What rights of first publication do the sequencers then have? As the discoverers of the sequence, they surely deserve some credit in the subsequent elucidation of function — credit that should extend beyond a simple reference to the database website. Yet, once someone else has annotated their sequence data, the sequencers' own ambitions in that direction have effectively been pre-empted.

These genomes are not mere scientific curiosities. For example, the sequences of the 14 chromosomes of the protozoan *Plasmodium falciparum* and the 11 of *Trypanosoma brucei* represent keys to significant advances in the treatment of malaria and sleeping sickness,

respectively. It would be self-evidently undesirable to allow the accumulating data, and those of other pathogens, to sit uninterpreted on open databases while sequencers concentrate on reaching 99.9 per cent completeness in projects that typically extend over years.

One could argue that those involved in sequencing cannot have both prompt openness and self-protection and that, once the data are publicly available, it is open season. Others might propose the opposite extreme: that the sequencers have the right even to co-authorship on papers that spring directly from their efforts. That would seem to amount to an extension of publishing practice in biology to recognize the fact that distinct but interdependent roles — in effect, collaboration — are played by those who produce basic information and those who accomplish interpretation.

Pragmatic solutions

Such relationships are already commonplace in the physical sciences: for example, in astronomy, the developer of a new detector is sometimes, by agreement, included on initial papers that emerge from the application of that device. In high-energy physics, huge collaborations are necessary, and everyone gets due credit by a listing of the hundreds involved. While these particular examples may not translate directly to biology, they are suggestive of a way forward.

That pragmatism seems further than some are prepared to go (see, for example, *Nature* 405, 601; 2000). Some sequencers firmly resist the idea that others can 'cream off' the data they so generously display. Some argue that people are rushing to annotate prematurely, and that waiting for more extensive and accurate sequence coverage is often advisable. It is also true that sequencing is sometimes perceived as scientifically narrow — some senior biologists, in response to media hype about genomics, have emphasized the scientific limitations of a genome sequence, highlighting all the work needed to reveal its meaning. It must be recognized that the scientists involved have dedicated themselves to a systematic project at some expense of their freedom to look more deeply at what they are uncovering. Perhaps not surprisingly, they may resist those whom they perceive to be taking such opportunities away from them.

On the other hand, researchers outside the sequencing collaborations who wish to annotate those data feel not only that the world should not have to wait, but also that the results of annotation should get maximum exposure through publication. Simply adding annotations to a database doesn't do justice to the significance of that work, nor to the need to make that insight available to the biomedical community in an intelligible and convenient form.

There is no simple or generic answer to such issues. Looking at them in a positive light, one can be thankful that there is so much pressure to reveal both the sequence and the function of these genomes as promptly as possible. But more time on annotation by sequencing researchers, and more collaborative goodwill between them and others seeking to use their sequences, would seem to be in order. ■





Hung out to dry
Still no European
money for biology
infrastructure
p723



Gene therapy
Regulation of trials
in question on both
sides of the Atlantic
p724



Climate change
US assesses impact of
global warming and
countermeasures
p725



Fast track
Chandra spots
outbursts from
neutron star
p727

Relations thaw between genome rivals as finish line draws near

Paul Smaglik, Washington

The war of words that has marked — many would say marred — progress towards sequencing the human genome seems to have reached a cease-fire. The move comes as both the publicly funded Human Genome Project (HGP) and its commercial rival, Celera Genomics of Rockville, Maryland, prepare to announce within the next few weeks that they have separately sequenced and assembled versions of the human genome.

The détente was orchestrated at a cancer conference at the National Institutes of Health in Bethesda, Maryland, last week. It comes months after negotiations between the two broke amid a bitter exchange of accusations (see *Nature* **404**, 117; 2000).

Consecutive speeches by Francis Collins, director of the National Center for Human Genome Research and leader of the public project, and Craig Venter, head of Celera, rekindled speculation that the projects may yet cooperate — or at least share the credit.

"There's no reason for there to be wars," said Venter, after Collins had reiterated that the projects are "complementary". Each praised the other's approach, where before they had pointed out perceived weaknesses.

But hints of the old chill slipped through. Collins mentioned repeatedly the benefits of the HGP releasing its information every 24 hours — a sticking point that has prevented



Jaw not war: in recent speeches, Collins (left) and Venter praised each other's approaches.

closer cooperation between the two projects. And Venter noted wryly that Collins' chronology of genomic milestones neglected to mention *Haemophilus influenzae* — the first bacterial genome, sequenced under Venter's leadership at The Institute for Genomic Research, in Rockville, Maryland.

In general, however, the tone was one of mutual praise. Collins, for example, lauded Venter's use of 'whole-genome shotgun' sequencing — in contrast to his earlier expressions of scepticism about its effectiveness for large genomes.

Shotgun sequencing involves breaking the genome into millions of fragments, read-

ing their ends, and then using computers to assemble them. Celera used the technique, along with gene maps built by Gerry Rubin, of the University of California at Berkeley, to assemble the fruitfly genome.

Collins is considering doing the same for the public project's version of the mouse genome. "We've learned a lot from Dr Venter and others about the quality of the whole-genome shotgun," he said.

Venter, in turn, seemed grateful for the public project's contribution to his work. One version of Celera's human genome will incorporate the HGP's data. "More data in any assembly will add more quality to the project," said Venter.

His remarks contrasted with previous criticism that the HGP's 'working draft' of the human sequence would probably contain too many 'holes'. In April, for example, Venter told a Congressional hearing that the HGP "may be at a stage where quality and scientific standards are sacrificed for credit".

Venter also applauded Collins' recent decision to sequence the mouse — which Celera is already about a third of the way through — and rat genomes.

James Watson, director of Cold Spring Harbor Laboratory, who has criticized Celera's plans to keep its information proprietary, noted that Celera had spurred the public project on. "From the public viewpoint, everyone has gained," he said.

China warned of AIDS 'disaster'

One of China's leading AIDS researchers warned last week that, if the country does not move to combat the disease, it could soon face a "national disaster", with the largest number of AIDS cases in the world.

According to the China News Agency, Zeng Yi, the director of the Institute of Demographical Research at the University of Beijing, told a meeting of the Chinese Academy of Sciences that AIDS threatened all the gains made since the country's economic reforms and opening up to the

West. But, he said, China has dragged its feet in its efforts to control the accelerating spread of HIV infection.

Zeng added that, among other things, there was a lack of investment in scientific research, and that the attitude of some officials was hampering prevention work. He told the academy that one researcher had been fired because local officials believed that, by revealing the AIDS situation in their region, he had made them look bad and hurt the region's image.

Canada unifies medical research community

David Spurgeon, Montreal

A new era opened for Canadian medical researchers last week with the launch of the Canadian Institutes of Health Research (CIHR). The new body replaces the Medical Research Council of Canada, which has been the major federal agency funding health research and training since 1960 (see *Nature* 393, 613; 1998).

The CIHR will bear some resemblance to the US National Institutes of Health. But there are also important differences. Whereas the NIH has 25 institutes and centres housed in 75 buildings in Bethesda, Maryland, the CIHR will have no main campus. Its researchers will be grouped into 'virtual' institutes, located in 16 health-science centres, 50 teaching hospitals and 65 research institutes across the country.

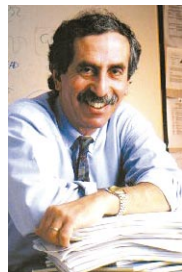
The aim is to link investigators in biomedical and clinical research, health systems and services research, and population-health research across Canada. There will be no intramural research in the NIH sense: instead, work will be done by investigators in the participating organizations, although it will be evaluated and reviewed centrally by peer committees.

Among the advantages — apart from a

much bigger budget — is said to be the inclusion of a broader range of health research than any previous agency, encouraging closer collaboration between disciplines and possibly new avenues of research.

Alan Bernstein, former director of the Samuel Lunenfeld Research Institute at Mount Sinai Hospital in Toronto, will be the CIHR's first president. A professor of molecular and medical genetics at the University of Toronto since 1984, Bernstein has made key contributions to the understanding of embryonic development, haematopoiesis, cancer and formation of the cardiovascular system.

Biomedical researchers, although initially apprehensive, have welcomed the new arrangements. Cecil Yip, vice-dean of research at the University of Toronto, says that the CIHR is "a good thing from many perspectives, not the least of which is that it represents a serious commitment [to health research] from the federal government".



Bernstein: will preside over 'virtual' institutes.

The budget of the institutes is set to double from the MRC's 1999–2000 figure of Can\$302.5 million (US\$184 million) over the next three years; it will be Can\$402 million in 2000–01, rising to Can\$533 million the following year. Henry Friesen, who was MRC president and chief architect of the CIHR, says that a budget of Can\$1 billion is both possible and defensible.

The creation of the CIHR was announced in the 1999 federal budget, following the report of a National Task Force on Health Research, made up of representatives from the whole medical research community.

A 19-member governing council will set the strategic direction, goals and policies of the component institutes — whose details have not yet been announced — and determine the mandate of each. It will also appoint the institutes' scientific directors and advisory boards.

Federal health minister Allan Rock says that the CIHR "represents an entirely new way of conducting health research ... its emphasis on partnerships with the voluntary healthcare sector, other government agencies and industry will be a model to be emulated around the world".

http://www.cihr.ca

Miniature antennas will eavesdrop on the Universe

Steve Nadis

Researchers at Ohio State University (OSU) are building a radio array, composed of 64 tiny antennas, as a prototype for telescopes that may revolutionize the search for extraterrestrial intelligence (SETI) and radioastronomy in general.

At roughly 100 square centimetres, each antenna is so small that its field of view extends from horizon to horizon, unlike dish-shaped antennas, which focus on a small portion of the sky. Eight antennas have been installed, with four already operating full-time. The team expects the entire array to be in place this autumn.

A telescope built this way could monitor the whole sky continuously over a wide range of radio frequencies. Not only might it pick up transient signals from extraterrestrial life, but it might also spot short-lived astrophysical phenomena missed by other instruments.

The project, funded by the California-based SETI Institute, has so far cost less than \$200,000. "We're working with a 64-element array because that's large enough to get into some of the technical issues, but small enough that we can put it together at a reasonable cost," says the principal

investigator Steven Ellingson, an electrical engineer at OSU. "Our main goal is to figure out how to build and operate an entirely new kind of telescope."

One lesson that has already been learned, says Ellingson, is that, rather than trying to form images of the entire sky, "it's more efficient to use modern signal-processing techniques to figure out when and where something interesting is happening, and then focus on just that part of the sky".

The prototype is called Argus, after a Greek mythological hero with 100 eyes. The urban setting of Columbus, Ohio, is far from ideal for astronomy, which is why the array is intended primarily as an engineering test-bed, rather than a front-line observational tool.

Nevertheless, Argus can still make valuable contributions to radio astronomy. The instrument could, for instance, detect bright, ephemeral events such as supernovae and advise other observatories on where to point their telescopes.

X-ray and gamma-ray bursts are another good example, says Frank Briggs, a radio astronomer at the University of Groningen in the Netherlands. "We can learn new things about these phenomena by looking at



the radio signals they emit, as well as at the X-rays and gamma rays."

Such an array could have tremendous potential, says Kent Cullers, a physicist at the SETI Institute — provided that costs can be kept down. The main expense for a larger, more sensitive instrument, perhaps linking a million antennas, would be the computing power tying them together, rather than the antennas themselves. "If computing costs keep coming down, this could well be the telescope of the future," says Cullers.



Hard up: the European Bioinformatics Institute outside Cambridge faces a precarious future.

European centres rebuffed in infrastructure funding bid

Quirin Schiermeier, Munich

Hopes that this week's meeting of Europe's research ministers might produce a quick political solution to the financial crisis threatening European life-sciences infrastructures appear to have been dashed.

The ministers have forced José Mariano Gago, Portugal's science minister and president of the European research council, to back down from a proposal to instruct the European Union (EU) to revise its existing funding rules, which exclude support of running costs for scientific facilities.

This is particularly disappointing for two such high-profile facilities: the European Bioinformatics Institute (EBI) at Hinxton Hall outside Cambridge and the European Mouse Mutant Archive (EMMA) at Monterotondo near Rome.

Although both were set up with support from the European Commission, funding from Brussels dried up last year as a result of a rule change in the EU's Fifth Framework Programme of Research (FP5). This now states that construction and operation of infrastructure in the life sciences are excluded from funding (see *Nature* **402**, 4; 1999).

Both institutions are being kept alive by rescue funds, from the Italian research council for EMMA, and the European Molecular Biology Laboratory for the EBI. Long-term security is lacking, even though Gago asked European research commissioner Philippe Busquin in spring to make funding available to the facilities (see *Nature* **404**, 217; 2000).

Gago also prepared a draft resolution to be presented to his fellow research ministers. But the initial draft, under which EMMA and EBI would have been re-eligible for funding under FP5, has been toned down significantly after member states expressed reservations. The revised draft merely "invites" the European Commission "to find a mechanism...

enabling the Community to find a solution" to the problems of funding EMMA and EBI.

Even so, several EU member states — as well as the commission itself — still want all reference to EMMA and EBI to be omitted from the resolution. They suggest that infrastructure funding should be addressed in the 6th Framework Programme, for which discussions start in September.

A commission official points out that the work programme in FP5's 'Quality of Life' Special Programme was modified in November by restricting the clause excluding infrastructure funding to "routine operations". This excludes costs for computation, or for freezing and storage of mouse embryos.

Busquin says the results of the last call for proposals should be monitored to assess the impact of the modified exclusion clause before the commission considers further modifications. But Gago says European scientists would then be "hostage to bureaucracy".

Meanwhile, scientists at EMMA and EBI remain in a state of uncertainty. "We are still in a position of serious crisis," says Graham Cameron, co-head of the EBI. The EBI's largest task, according to Cameron, is the setting up of 'ArrayExpress', a public-domain repository for microarray gene-sequence expression data, which the EBI is pursuing with the help of industry while still seeking long-term government support.

EBI scientists have applied for funding from FP5 for projects worth 30 million euros (US\$29 million), of which only 1.5 million euros has so far been granted. EMMA gave up applying for FP5 money after its 1999 applications were turned down.

Infrastructure funding is a priority issue for the proposed 'European Research Area', a new initiative from Busquin. Options will be discussed in September at a high-level EU conference in Strasbourg. ■

Spanish postgrads push for better employment rights

Xavier Bosch, Barcelona

Postgraduate students in Spain have set up a nationwide organization in an attempt to improve their working conditions.

Known as the 'Precarious National Federation' — to reflect the situation in which the students see themselves — the new organization wants to see the current system of monthly student grants replaced by 'training contracts' that last for the whole length of the students' studies. The federation also wants all postgraduates to be given the status of 'research personnel in training'.

The creation of the federation reflects widespread dissatisfaction among students over their employment situation. Manuel Pérez-Mendoza, president of the federation and a research chemist at the University of Granada, says most postgraduates in Spain have no employment rights, such as affiliation to the social security system or health benefits.

Elisenda Vendrell, co-founder of the federation and a biologist at the Cancer Research Institute in Barcelona, argues that the work of postgraduate students "is seriously undervalued in the labour market".

She points out that even though the students spend a long time doing grant-supported research, as well as teaching, their work is not properly recognized. "For example, we have no representation in the governing bodies of public research institutions," she says.

Researchers in the federation have been making their case to senior university officials and those responsible for science in the main political parties. They also intend to negotiate directly with the Ministry of Education and Culture and the recently launched Ministry of Science and Technology.

The postgraduates have received support from some senior researchers. "Governments tend to adopt a rather cynical attitude towards young researchers, because they know that their eagerness to do research can lead them to accepting working conditions which would be unacceptable at a comparable level in other fields of employment," says John Beckman, research professor at the Higher Council of Scientific Research's Astrophysics Institute of the Canary Islands. "This situation is accentuated in Spain because of the rigidity of its administrative system". ■

NIH panel may increase gene-trial scrutiny...

Paul Smaglik, Washington

The US National Institutes of Health (NIH) was last week urged to give its Recombinant DNA Advisory Committee (RAC) a greater role in the regulation of clinical trials for gene therapy. The call came from a working group set up to advise the NIH in the face of increasing concern over the conduct of such trials.

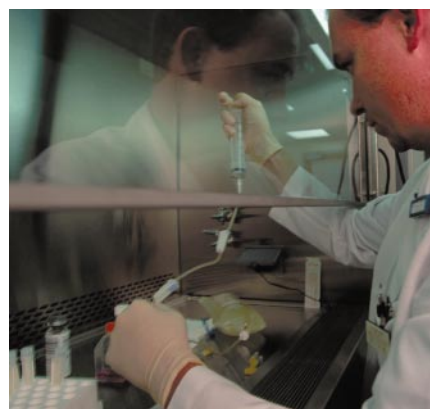
The advice runs counter to the policy of the former NIH administration, which had sought to diminish the RAC's involvement. It is also contrary to the wishes of clinical investigators, who are concerned about increasing amounts of red tape. But with legislators pushing for tighter control on clinical trials, such a move could reduce the pressure for congressionally mandated regulation.

Under the new proposals, enrolment in a clinical trial would not begin until the RAC

had accepted the trial's protocols. If the committee found that a protocol was novel — because, for example, it used a new kind of vector to deliver a gene, or sought to treat a new group of patients — the trial's investigator would have to present it formally to the RAC, which meets four times a year. Only after the committee was satisfied with the protocol would enrolment begin.

Christine Cassel of Mount Sinai Medical Center, and co-chair of the working group, told the NIH director's advisory committee meeting last week that only about 10 per cent of protocols would be subjected to this higher level of scrutiny. Before the new guidelines can come into effect, they need to be endorsed by acting NIH director Ruth Kirschstein, who has asked her advisory committee to comment on the proposals.

Claudia Mickelson, biosafety officer at



PAUL A. SOUDERS/CORBIS

Bench-top bothers: there are fears that tighter RAC review could slow down clinical trials.

the Massachusetts Institute of Technology, RAC chair and a member of the working group that drafted the recommendations, denies that the RAC review would slow down the clinical-trial process. The proposals call for simultaneous review by the investigator's institutional review board, institutional biosafety committee and the RAC. Until now, such processes have been conducted sequentially.

In the past, novel protocols were sometimes reviewed by the RAC at the same time that investigators were enrolling patients, or even proceeding with the trials. "That's an untenable situation," says LeRoy Walters, former RAC chair and director of the Kennedy Institute of Ethics at Georgetown University in Washington DC, who has long been calling for the body to have more regulatory authority. "That just promotes disrespect."

Walters also notes that the new guidelines emphasize that the RAC is an advisory body to both the NIH and the Food and Drug Administration (FDA). These two bodies have sometimes had difficulty communicating with each other — especially over adverse events in clinical trials. This is another area the working group hopes to address, by harmonizing reporting to both agencies.

Even if the FDA recognizes the RAC as an advisory body, Walters says he has a "nagging concern" that gene-therapy trials funded and carried out by pharmaceutical companies could never have their adverse events made public. The RAC aims to disseminate such information in order to detect patterns and prevent the use of dangerous vectors, but the pharmaceutical industry is keen to keep adverse events confidential.

"We don't know that there aren't human gene-transfer protocols that have been submitted to the FDA that are totally invisible to the public review process," says Walters.

...as Europe's 'excessive secrecy' is deplored

Declan Butler, Paris

The recent difficulties faced by gene-therapy trials in the United States seem to have had little impact on attitudes to the regulation of such trials on the other side of the Atlantic.

This was the prevailing sentiment at a meeting of European scientists, regulators and industrialists involved in gene therapy, held in Paris last weekend. Indeed, several delegates deplored what they described as the excessive secrecy of European approval procedures for gene therapy.

The meeting was organized by Euregenethy, a grassroots network of scientists and physicians keen to promote gene therapy and the harmonization of regulatory laws across Europe.

The current plethora of national committees and rules is widely seen as a hindrance to clinical development, providing less clarity and concentration of expertise than would a single pan-European procedure, and making it more difficult to organize the large trials needed to assess safety.

Speaker after speaker told the meeting that Europe's handling of gene therapy is anything but harmonized. Although new drugs are approved by the London-based European Medicines Evaluation

Agency, national governments must approve clinical trials. Only the United Kingdom, France, the Netherlands and Italy have dedicated systems for gene therapy in place.

In all countries, approvals depend on several different committees. Mark Richardson, head of regulatory affairs at Orio clinical services in Slough, UK, said that in such multi-stage reviews, the division of responsibilities between scientific, ethical and regulatory bodies is often unclear. As a result, approvals face long delays, and criteria for approval are unclear and differ from country to country.

The main organizer of the meeting, Odile Cohen-Haguenaer of the St Louis Hospital in Paris, is working on gene therapy for Fanconi anaemia, a rare childhood disorder. Despite progress in the laboratory, a multinational clinical trial will be "extremely difficult" to set up in Europe, she says.

Ted Friedmann, director of the programme in human gene therapy at the University of California, San Diego, and a member of the US Recombinant DNA Advisory Committee, said the meeting had taught him the "political realities" of the complexity of European regulation. Whatever the recent failings of the

US system, he said, it was at least "linear", and gave "much less opportunity for confusion".

Friedmann also said that if there was a lesson for Europe from the US controversy, it was of the need for building "openness" into regulatory structures: "over the last year we have learnt that open review is better than closed" (see *Nature* 405, 606–607; 2000). A marriage had to be found between legitimate industrial confidentiality "and the need for the public to know what happens".

Following the death of Jesse Gelsinger during a clinical trial run by the University of Pennsylvania's Institute for Human Gene Therapy in which high doses of adenoviral vector particles were given, Euregenethy has been carrying out a survey on the use of such vectors in Europe. It has found that many of the 30 ongoing or finished clinical trials have used them, some in similarly high doses.

Trials of the systemic application of adenoviral vectors have been put on hold by scientists themselves following Gelsinger's death, according to Euregenethy, while various European countries have begun to assess the safety of trials that have used adenoviral vectors.

◆ <http://193.48.40.240/www/euregenethy/euregenethy.html>

UK researchers call for limits on animal experiment 'red tape'

Natasha Loder, London

More than 100 leading UK medical researchers have written to British science minister David Sainsbury to ask for an end to what they claim are excessive delays and red tape involved in regulating animal experimentation.

The researchers, who include five winners of the Nobel Prize for Medicine and 38 Fellows of the Royal Society, say they are worried that Britain could lose its leading position in biomedical science.

"It can now take six months or longer to obtain approval for a research project using animals in the UK, whilst in other countries it can take weeks or even days," the authors of the letter write. They emphasize that animal studies are likely to increase in importance once the human genome sequence is completed.

They also report that some researchers are now taking their work abroad in response to the British situation. One researcher says that he now does some of his work in the United States because of the "tardiness of the Home Office in issuing appropriate licenses".

One of the signatories to the letter is Nancy Rothwell, professor of physiology at the University of Manchester. She is also a member of the UK Life Sciences Committee (UKLSC), which promotes the interests of a number of scientific societies. Rothwell says that the UKLSC is keen to start a dialogue with the Home Office to see how the current situation might be improved. But there will be no simple answer, she says.

Bob Combes, scientific director of the Fund for the Replacement of Animals in Medical Experiments, says: "If there is good science being done and it has to be done with animals then sooner done the better. We fully support the removal of unnecessary bureaucracy, but the question is: what is [unnecessary]?"

Combes says that many scientists see the checks and balances in the current regulations as bureaucratic. "But we see them as necessary. Unfortunately, not all scientists appreciate that animal welfare should be inextricably linked to best scientific practice."

Many of the UKLSC's member societies feel that the UK government has done little to listen to their worries over the regulations. But Sainsbury said this week that he would look into the scientists' concerns, and has agreed to meet with a representative delegation of the letter's signatories. ■

US climate report underlines local impacts of warming

Paul Smaglik, Washington

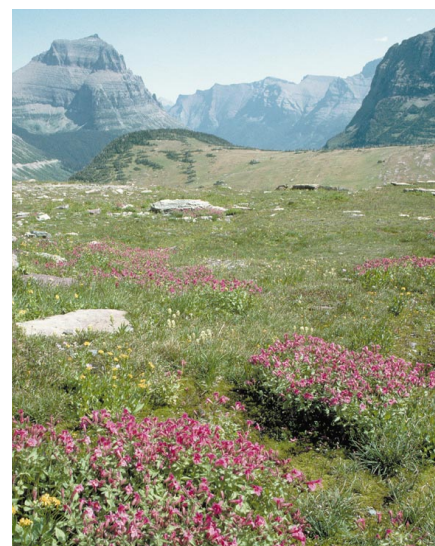
Americans are finding out this week how radically a change in climate could affect their lives. The US Global Change Research Program (GCRP) has issued a draft report outlining some of the dramatic effects global warming could have on the country.

If, the report predicts, the average temperature of the United States rises by five to ten degrees Fahrenheit over the next 100 years, the Great Lakes and the southwestern desert would shrink significantly, and rates of thawing would increase in Alaska.

The GCRP was set up by Congress in 1990, and brings together the efforts of ten federal departments and agencies, as well as outside organizations. The report has been under preparation since 1997, when it was started by a series of 20 workshops held around the country to identify the critical interfaces between climate change, the environment, and society.

Its draft conclusions focus on the probable consequences of global warming, rather than on how this will occur. But that approach, while dramatizing the potential impact, opens the conclusions to criticism, as regional climate modelling is on less firm scientific ground than global modelling.

Michael MacCracken, director of the National Assessment Coordination Office of the GCRP, says that the projections are all "plausible" and are scored by likelihood, based on the best available computer modelling. But they include some 'wiggle room'. For example, the Southeast will become either hot and dry or warm and moist—but not hot and moist. The report stresses that, given the complexity of the environment and its interaction with climate, both surprises and uncertainties are likely.



Rocky future? Global warming could mean bad news for alpine meadows.

Environmental groups have applauded the report for pointing out that natural ecosystems, such as alpine meadows in the Rocky Mountains, could be especially vulnerable to global warming. "The assessment shows that many of the country's distinct natural features could deteriorate as a result of changing climate," says Susan Subak, a senior research associate at the Natural Resources Defense Council.

Industry groups are less pleased. "We're troubled by the pessimistic nature of some of the things that it says, especially when the computer models are very contradictory," says Frank Maisano, spokesman for the Global Climate Coalition. Maisano notes that regional predictions of climate change are unreliable. ■

♦ <http://www.gcrao.org/NationalAssessment/sgrs.html>

Los Alamos 'loses' key weapons data

Rex Dalton, San Diego

The Los Alamos National Laboratory in New Mexico is facing new embarrassment over its security procedures this week, after admitting that it has lost computer hard drives containing classified information about the design of nuclear weapons.

The drives were found to be missing when the laboratory was being cleared of important equipment as brush fires approached last month (see *Nature* 405, 264; 2000). "This is an extremely serious matter and we are taking swift actions to deal with it," laboratory director John Browne said on Monday.

The laboratory is run by the University of California on behalf of the Department of Energy. Browne said that a joint investigation is being carried out by the department and the Federal Bureau of Investigation. He has promised "appropriate disciplinary actions" if individuals are found to be responsible.

The missing data are believed to include information about how to defuse nuclear devices in case of emergencies, as well as data on Russian weapons. A statement from the laboratory said that efforts were continuing to locate the missing media "or to determine if they were inadvertently destroyed". ■

Celera and Geron join forces over stem cell gene analysis

Washington The sequencing company Celera Genomics and Geron Corporation — which specializes in the biology of human stem cells — announced this week that they are to collaborate on the identification and functional analysis of genes involved in human cell differentiation and early human development.

Geron will develop and commercialize a number of small-molecule drugs, protein therapeutic products, cell and gene therapies, and prenatal diagnostics. Celera, which is about to complete its sequence of the human genome (see page 721), will use the information to enhance the annotation of the genome and develop and commercialize probe sets for the analysis of gene expression.

“We will be using human pluripotent stem cells — the most basic form of human cells that contain a diverse set of genes not expressed in high abundance in other cells — as a source to better understand the human genome,” Craig Venter, Celera’s president and chief scientific officer, said in a statement.

Thomas Okarma, Geron’s president and chief executive, said that the project “will bring us one step closer to realizing

the therapeutic potential of human pluripotent stem cells”.

US environment agency urged to bolster research

Washington The US Environmental Protection Agency should strengthen its research base, according to a report from the National Research Council (NRC) published this week. The report says that stronger science would help the agency to direct its resources: “We need scientific information to avoid wastefully targeting inconsequential problems while ignoring greater risks.”

The NRC advises the agency to create a new position, deputy administrator for science and technology, with agency-wide responsibility for science policy. Creating such a position would require authorization from Congress, appointment by the president, and confirmation by the Senate.

CNRS joins in on muon neutrino beamline

Paris France’s Centre National de la Recherche Scientifique (CNRS) has announced that it will contribute to the construction of a neutrino beamline being built under the Alps from Geneva to the Gran Sasso laboratories near Rome, along

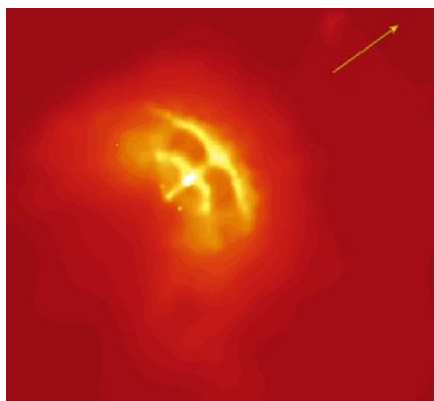
with experiments scheduled to begin in 2005. Four laboratories of the CNRS and the National Institute for Nuclear and Particle Physics will participate in building a beamline that will carry muon neutrinos on a 730-km journey.

Researchers from 20 institutes around the world are collaborating on the project, coordinated by Italy’s National Institute for Nuclear Physics (see *Nature* **402**, 847; 1999). The experiments are designed to measure the creation of tau neutrinos.

Drosophila find new home in Tucson, Arizona

San Diego The University of Arizona at Tucson is to be the new home of the US National *Drosophila* Species Stock Center, which provides more than 300 types of fruitfly to scientists worldwide. The National Science Foundation (NSF) recently awarded Arizona the contract to maintain the living fly archive, which has until now been held at Bowling Green State University in Ohio.

The NSF will pay \$1 million to relocate the fly archive, which stocks 1,500 cultures for research in genetics, physiology, neurobiology and ecology. It will be housed in Arizona’s Center for Insect Science.



Secrets of a neutron star shine through

Washington Intriguing new pictures from the Chandra X-Ray Observatory (above) show a neutron star ejecting dramatic rings and jets of high-energy particles while moving through space. The neutron star is the brightest spot in the picture, to the left of centre. Jets are being emitted from the north and south poles in the direction of the top-right and bottom-left corners of the picture. The rings of gas coming from the equator of the dense and rapidly rotating star are seen at about 45 degrees to the line of sight.

What is exciting scientists is that the jets seem to point in the direction of motion. For

the star to move it had to be given a kick, which could have been either a non-symmetric explosion or the results of the jets acting like exhausts, with a more powerful one acting in a backward direction.

Office for Human Research Protections gets director

Washington Greg Koski has been appointed the first director of the Office for Human Research Protections (OHRP), a new office at the US Department of Health and Human Services (DHHS). The office replaces the Office of Protection from Research Risks run by the National Institutes of Health (NIH).

The advisory committee to the director of the NIH recommended last year that the new office be located outside the NIH, directly under the DHHS secretary (see *Nature* 399, 514; 1999). Koski is currently the director of human research affairs at Partners HealthCare System in Boston, and associate professor of anaesthesia at Harvard Medical School.

Royal Society sets up study of GM animals

London Britain's Royal Society is to carry out a study into the use of genetically modified (GM) animals in medicine and agriculture. Eight fellows of the society will review

scientific progress and advise on future policy implications. They will also look at likely future research, and consider legislation governing the use of GM animals.

The review is a response to public concern over GM animals and the potential risks posed to humans and the environment. The group is seeking evidence from interested parties. The main medical interest in GM animals is in creating animal models for human disease.

Mars Polar Lander returns to haunt NASA

Washington The US space agency NASA has attracted the ridicule of Representative James Sensenbrenner (Republican, Wisconsin), chair of the House Science Committee, for declaring the robotic arm of the ill-fated Mars Polar Lander to have "achieved" its performance target before the launch of the mission, which was lost last December (see *Nature* 402, 565; 1999).

The description was used in the space agency's Performance Report for 1999, and was based on the arm passing its acceptance tests. "If this is NASA's definition of success, planetary exploration just got a whole lot easier," Sensenbrenner said in a statement. A NASA spokeswoman has been quoted as saying that the wording "must be an error".

One for all — and all for one

Hubert Markl, president of Germany's Max Planck Society, wants to make the organization work as a coherent whole. As the society prepared for its annual meeting in Munich, he explained his vision to Alison Abbott.

Four years into his six-year term as head of the Max Planck Society (MPS), zoologist Hubert Markl is shaping his legacy. If all goes to plan, his presidency will be remembered for turning the MPS into a more dynamic organization, able to react quickly to the opportunities offered by new avenues in research. And last week's publication of *Research Perspectives 2000 +*, a blueprint for the society's future, may well be seen as the turning point.

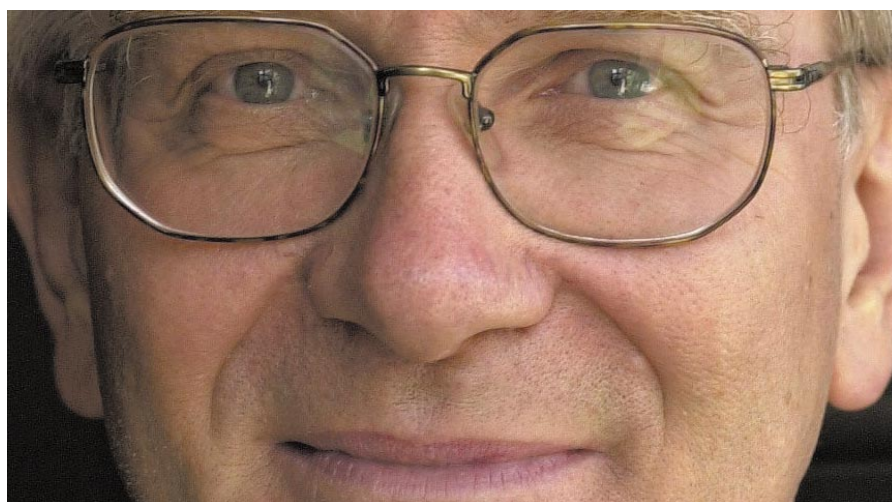
The document itself is unremarkable — it identifies probable future research directions, listing them under broad themes such as "From genes to organisms". But it is the process that produced *Research Perspectives 2000 +* that Markl sees as the catalyst for change. In particular, he believes it will now be easier to close existing Max Planck Institutes so that new ones can be born — a perennial problem for MPS presidents wanting the flexibility to respond to new research opportunities.

Closing time

No one questions the society's reputation for scientific excellence — it has nurtured 15 Nobel prizewinners since the Second World War. But if the MPS has a problem, it is a lack of flexibility. Arguably, both the society's strengths and its weaknesses result from the established principle that Max Planck Institutes are built around their directors. The selection of a new director is a long and carefully considered process. But once installed, the director is left to pursue whatever research he or she thinks fit.

When a director retires, the MPS reviews the work of the institute concerned and, if appropriate, appoints a new director with different research interests. The institute may then be closed and reincarnated in another guise. But in practice, as research directors come up to retirement, they nearly always fight against closure as they feel it would undermine their lifetime's achievement. Over the years, the protests have been sufficient to frustrate the society's attempts to remodel several of its institutes.

As an expert in animal behaviour, Markl knows all about the strength of the biological urge to reproduce. But from experience with retiring Max Planck directors, he has noticed that "the cultural reproductive urge is stronger". Markl had to close down six institutes and more than 20 departments within individual institutes in western Germany to



DPA/FRANK MÄCHLER

This gives everyone the chance to make their point about how they see their future

Hubert Markl.

help finance the expansion eastwards that followed German unification. "It is not easy to make changes," he says. "There were many who considered me absolutely heartless."

Markl believes the consultative process used to draft *Research Perspectives 2000 +* will make such changes easier. Each of the 78 Max Planck Institutes was asked to submit its plans for the future. These were then debated widely and synthesized into the final document by a committee drawn from the research directors who sit on the society's governing senate. For the first time, the MPS's directors were forced to think strategically as a group, regardless of disciplinary boundaries.

The result, says Markl, is a vision for the future that everyone, in theory, has signed up to — making it difficult to defend the retention of institutes that have not presented convincing plans. The process will be repeated every few years, to keep the society's ideas fresh. "This new instrument gives everyone the chance to make their point about how they see their future," says Markl.

Markl has also introduced a parallel system of evaluating entire disciplines, in which

MPS research at relevant institutes is compared with developments elsewhere in the world. The first evaluation, on molecular neuroscience, has just been completed.

Towards a flexible future

Research Perspectives 2000 + was presented to the research directors last week, at the society's annual meeting. Although most have yet to digest its contents, the document has been received warmly by senior figures outside the MPS. Josef Lange, state secretary for science, research and culture in the city of Berlin, thinks it will also influence Germany's universities, particularly if another of Markl's innovations — a new system of graduate research schools — works as planned. Nine MPS International Research Schools, offering PhD training in conjunction with universities, will open this autumn. Markl believes they will stimulate collaboration between the MPS and the university sector, which is not as well developed as many policymakers would like. He hopes that there will be 30 of these schools within five years.

But has Markl really found a system that will give the MPS the flexibility to close institutes without getting bogged down in heart-wrenching protests? Lange is unsure. "It is a big task to come to a system of establishing and closing institutes," he says. The answer should become clear over the next few years, as a large number of directors who were hired during the society's expansion in the 1960s and early 1970s reach retirement age.

Alison Abbott is *Nature's* senior European correspondent.

Nanotech thinks big

The science of the incredibly small is shedding its sci-fi image. An anticipated influx of US government funds is nurturing a new wave of interdisciplinary nanoscale research, says Colin Macilwain.

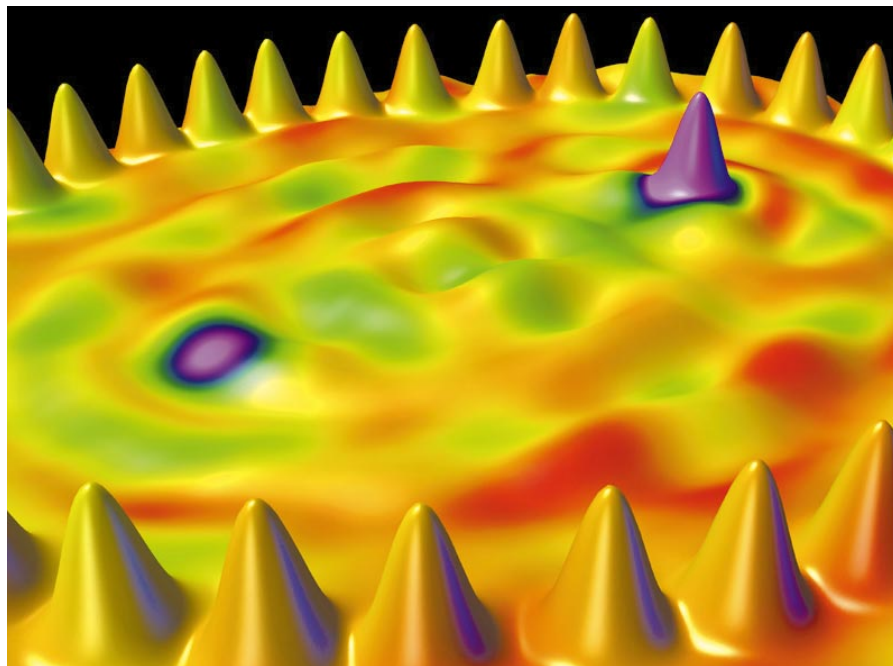
There is no such thing as bad publicity, it is sometimes said. But ever since Eric Drexler brought the term 'nanotechnology' into vogue in his 1986 book *Engines of Creation*, some researchers have felt that the field has been burdened by unwanted baggage. Drexler envisioned an era in which factory production lines were replaced by self-replicating, nanoscale 'assemblers' — and warned that such entities could supplant humans to become the dominant 'life' forms on our planet. These ideas were quickly seized on by transhumanists — people who imagine what the world will look like after technology has rendered us extinct.

But today, nanotechnology is acquiring the respect researchers in the field believe it deserves. This year, for the first time, the US government has identified nanoscale science and technology as a top research priority. The European Union and other nations are contemplating similar action, and bright young scientists and engineers are starting to gravitate towards the field. "The intellectual drive is even more important than the funds at the moment," says Mihail Roco, nanotechnology programme manager at the US National Science Foundation (NSF), and one of the prime movers in pushing the field up the federal research agenda.

Back to basics

The US Congress has been asked to fund a new National Nanotechnology Initiative, which would double federal funding for the discipline to \$500 million in 2001. When President Bill Clinton launched the initiative, his rhetoric might have made some researchers wince — speaking at the California Institute of Technology on 21 January, he boasted of "shrinking the information housed in the Library of Congress into a device the size of a sugar cube". But in reality, the initiative is not built on such grandiose promises. Its core element is a thorough programme of basic research — totalling \$170 million in the 2001 budget request — to investigate the behaviour of materials at scales ranging from 1 to 100 nanometres (a copper atom, for comparison, measures 0.25 nanometres across). This should provide the foundations on which tomorrow's nanotechnology will be built.

Increased effort in basic research is a priority, according to researchers in the field. One of the most exciting areas is the study of



Atomic insight: the scanning tunnelling microscope gives nanoscopic views, such as this quantum well.

carbon nanotubes — sheets of carbon rolled up into tiny, hollow cylinders. These can have alluring semiconducting properties, and so could be used to make nanoscale electronic devices. But the tubes remain awkward to manipulate and are difficult to 'grow' consistently. What is more, exactly how the tubes acquire semiconducting properties is still debated — the property could, for instance, be influenced by some unidentified impurity. Since the discovery of nanotubes, "there's been an explosive growth in papers, but not in understanding", laments Phaeton Avouris, whose team at IBM's Thomas J. Watson research centre at Yorktown Heights, New York, is trying to find some of the answers.

As well as boosting basic research, the new initiative promises to forge links across traditional disciplinary boundaries. Already, solid-state physicists, chemists, engineers and — most recently — biologists are uniting into rapidly growing nanoscale research groups at many of the leading US universities. Nanoscience, says Roco, is proving to be "an unexpected meeting point" for scientists from diverse backgrounds. Interdisciplinary collaborations often have to be forced by funding agencies, he says. But in nanoscience, they seem to be forming naturally as researchers from different areas share the same tools.

Nanotechnology gained an early impetus from the invention of what is still an important tool: the scanning tunnelling microscope (STM). In 1982, Gerd Binnig and Heinrich Rohrer at IBM's Zürich Research Laboratory made a discovery that would earn them the 1986 Nobel Prize for Physics. They made electrons 'tunnel' through a narrow gap from a pointed probe to an atomic surface, and showed that the strength of the resulting current is extremely sensitive to the distance between probe and surface. This enables the STM to map surface topography at a resolution of a few tenths of a nanometre. Researchers soon realized that the device could not only image individual atoms, but could also move them around one at a time. In 1991, a team at IBM famously inscribed the company's logo on a nickel surface using 35 atoms of xenon.

Playing atomic marbles

The STM was joined in 1986 by the atomic force microscope (AFM). This versatile device, also invented by a team led by Binnig, uses a laser to measure the deflection of a cantilevered stylus caused by electrostatic, chemical or magnetic interactions between the tip of the stylus and an atomic surface. It can map the surfaces of solids under gaseous

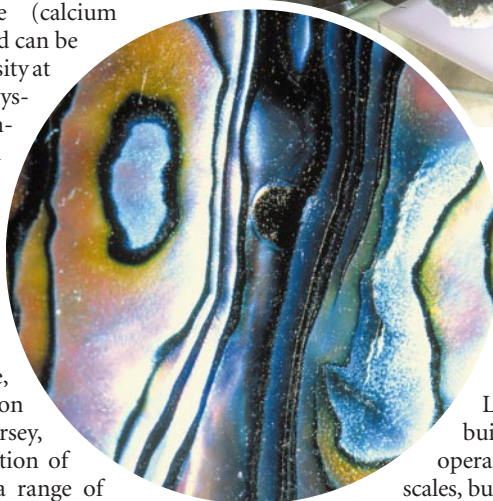
or liquid conditions, and can even produce images of biological molecules.

But the ability to observe atomic surfaces and to manipulate a few atoms at a time will not, by itself, lead to technological applications. "It's not about controlling things at the atomic scale — we do that already," explains Tom Theis, head of physical sciences at IBM's Watson centre.

One critical component is self-assembly — the design of molecular building blocks that will automatically link together to form nanostructures. Among the leaders in this area is George Whitesides at Harvard University. Last year, for example, he showed how molecules that self-assemble into a monolayer on a metal substrate can direct the formation of calcite (calcium carbonate) crystals, and can be used to control the density at which arrays of the crystals form¹. Such fine control over crystal growth is key to the fabrication of useful nanostructures, experts say.

Another important aspect is the integration of the nanoscale with larger structures. For example, İlhan Aksay at Princeton University in New Jersey, is directing the formation of 'colloidal crystals' at a range of length scales. The colloidal crystals are formed as tiny particles are deposited from suspension in a fluid onto the surface of an electrode. Four years ago, Aksay's team showed that the nanostructure of these crystals can be controlled by adjusting the strength or frequency of the applied electric field². Earlier this year, he showed how the simultaneous use of ultraviolet lithography can organize these crystals at the microscale³. "We are making materials with well-defined patterns not only at the nanoscale, but at longer length scales," says Aksay.

Living systems achieve this feat routinely, as Nobel laureate Richard Feynman noted in 1959 in a seminal lecture — called



Building blocks: by studying proteins used in the formation of abalone shells (left), Belcher (above) hopes to be able to control crystal growth.



"There is plenty of room at the bottom" — that foresaw the development of nanoscience.

Living things are built by processes that operate on the tiniest of scales, but which can create structures as large as whales and giant redwoods. Recognizing Feynman's prescience, nanoscientists are increasingly turning to biology for inspiration — although biological systems operate at spatial resolutions a billion or more times higher than the tiniest devices created by nanotechnologists to date. "That gives you an idea of how far we've got to go," says Theis.

Biologists get toolled up

Yet researchers working at the interface between the physical and life sciences are making rapid progress. Theis cites the "tremendously exciting" work of Angela Belcher of the University of Texas at Austin as an example. Belcher has established how relatively small quantities of proteins determine the structure of crystals of calcium carbonate in abalone shells⁴, making them 3,000 times stronger than the pure mineral.

After studying the properties of these proteins with AFMs⁵, her team has now gone on to identify proteins that bind in a similar way with semiconducting materials such as gallium arsenide and silicon — and published these findings in last week's *Nature*⁶. Such proteins, Belcher believes, could one day be used to direct the assembly of nanoscale electronic devices. "We are finding things that can bind with, and can

discriminate between, different crystallographic orientations, just like nature does," she says.

At IBM's Zürich centre, meanwhile, researchers led by Jim Gimzewski have constructed arrays of tiny, gold-plated silicon cantilevers. They hope that these structures will allow them to develop a new generation of chips that use nanomechanics, rather than radioactive or fluorescent tagging, to analyse DNA and other biomolecules. Already, the team has applied DNA bases to the tips of the cantilevers and shown that a simple array can distinguish between genetic sequences that differ by just one base⁷.

One problem with nanoscale biology is that biological samples are hard to manipulate — living cells tend to stick stubbornly to typical substrates such as silicon. But following the lead of Whitesides, who in the 1980s pioneered the use of ultraviolet lithography to cut moulds from silicon and so make tiny silicone-rubber components, researchers now use nanoengineered silicon templates to make components out of more bio-friendly materials. Today, they use electron-beam lithography to cut the moulds with even greater precision.

For example, Bob Austin, a Princeton biophysicist, is working to produce a 'lab-on-a-chip' that will manipulate and process living cells, extracting their DNA for analysis. Building on earlier work, in which his team showed that simple urethane-coated microarrays could be used to sort white blood cells into their various classes⁸, Austin is now working with layers of silicone rubber, to fashion structures that ultimately will include tiny valves and pumps. In essence, Austin aims to build a microscopic process plant. At this scale, working with fluids is very different from dealing with bulk flows,



Clinton: backs nanotechnology as a priority.

as the motion of the liquids is governed by the rules of capillary flow.

Nanoscience has also begun to attract the attention of the US National Institutes of Health, which will host a nanotechnology and biomedicine conference on 25–26 June. Last month, its National Cancer Institute (NCI) in Bethesda, Maryland, signed an agreement with the space agency NASA to study the medical potential of nanoparticles. Unlike particles in conventional powders, these nanoparticles have precisely defined and uniform dimensions. Such particles could be useful for drug delivery within the body. At present, administering drugs that do not dissolve in water is difficult. But nanoparticles could carry these molecules around the body, suspended in the blood.

Nanoparticles might also be used in diagnosis, bearing molecules that will bind to specific targets through the bloodstream — for instance, molecules on the surface of particular types of cancer cell. NASA will help the NCI with remote-sensing technologies to monitor the behaviour of such nanoparticles in the body.

Battling for nanoscopic mastery

Given this upsurge in interest, and the lure of funding from the nanotechnology initiative, universities across the United States are racing to stake their claims in nanoscale science. Stanford University in California and Cornell University in Ithaca, New York, are the two main 'nodes' in the existing network of nanotechnology centres established by the NSF. Both house major nanofabrication facilities, with clean rooms containing millions-of-dollars-worth of equipment — outside of industrial laboratories, only Delft University of Technology in the Netherlands boasts such an impressive range.

But Stanford and Cornell are about to get some competition. Earlier this year, Northwestern University in Evanston, Illinois, announced plans to spend \$30 million on a nanofabrication facility of its own. Harvard intends to make a similar investment on a new Center for Imaging and Mesoscale Structures — 'mesoscale' being a term long-used by physicists to refer to structures with dimensions measured in micrometres or nanometres.

Web links

National Nanotechnology Initiative

► <http://www.nano.gov>

Engines of Creation

► <http://www.foresight.org/E0C>

Stanford Nanofabrication Facility

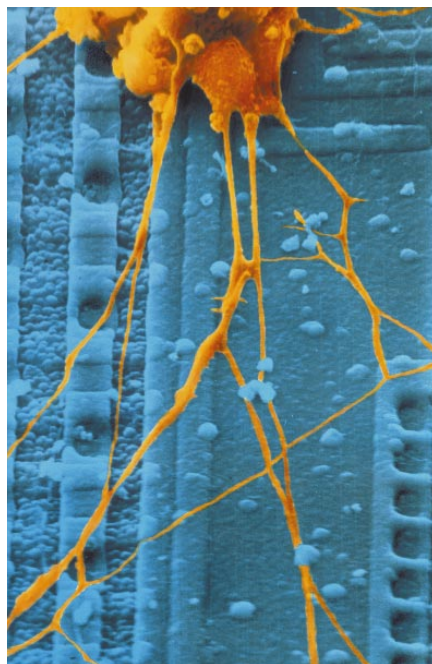
► <http://www-snf.stanford.edu>

Cornell Nanofabrication Facility

► <http://www.cnf.cornell.edu>

NIH conference on nanotechnology and biomedicine

► <http://www.masimax.com/becon/index.html>



Stuck up: cell samples adhere firmly to silicon, making them hard to handle on the nanoscale.

But the established centres are not sitting still. The Stanford Nanofabrication Facility is housed at the university's Center for Integrated Systems, which was built in 1985 with \$15 million of backing from 20 industrial sponsors. The facility pioneered the development of RISC (reduced instruction-set computer) silicon microprocessors, used today in many computers, but it is now moving aggressively into interdisciplinary research. "We're committed to opening doors to new academic disciplines," says John Shott, its associate director, who expects "exciting results" in areas ranging from biotechnology to integrated optics. "The electrical engineering department is hiring faculty whose main thrust is to collaborate with people in biomedicine and neuroscience," adds Fabian Pease, a professor in that department.

The Cornell Nanofabrication Facility expects to attract around 450 researchers this year — half of them visiting scientists. "The number of users is increasing by 15 to 20 per cent a year," says Sandip Tiwari, who directs the facility. Cornell plans to spend \$50 million on a new building for the facility, and has just won a \$20 million, five-year grant from the NSF to operate a new nanobiotechnology centre. This will be directed by Cornell nanoscientist Harold Craighead, and will make nanoscale tools available to biologists.

Few biologists have woken up to the potential of these tools, says Barbara Baird, a Cornell biophysical chemist who works on the production of biosensors. "But once it is demonstrated in one system there will be a rush from immunologists and everyone else," she predicts.

"The traditional barriers between fields are going away," agrees Richard Siegel, a



Cutting edge: the plasma etching tool at Cornell is part of the impressive nanotech facility.

pioneer in the use of nanoparticles, who is now head of materials science at Rensselaer Polytechnic Institute in Troy, New York. His department lies in the institute's engineering faculty, but Siegel is now pulling in young scientists from a variety of disciplines.

Inevitably, some are sceptical about this new interdisciplinary dawn. Funding for cross-disciplinary research is always fragile, sceptics note, and the researchers involved can get left out on a limb when projects end. "The biologists are only interested because the physicists are building new tools," says one physicist at Cornell. "If the tools aren't productive, they'll separate again."

But, at least for now, interdisciplinary nanotechnology has strong support from the NSF, and the prospect of wider backing from the entire US government, whatever happens in this year's presidential election. Although Congress is likely to trim back Clinton's \$500 million budget request somewhat, politicians of all hues with an interest in science keep stressing the importance of nanotechnology — which indicates that the field's current high profile is not likely to disappear overnight.

Despite this move to the mainstream, the transhumanist connection has not entirely been laid to rest. In the April issue of *Wired* magazine, for instance, Bill Joy, chief technology officer with Sun Microsystems, went as far as to argue that Drexler had been "naive" in underestimating the dangers posed by nanotechnology. But as the fruits of Clinton's National Nanotechnology Initiative begin to be harvested, its supporters predict that concerns about malevolent nanoassemblers will be replaced by excitement over the field's scientific and economic potential. "We are laying the foundations for the next industrial revolution," declares Theis.

Colin Macilwain is *Nature's* Senior US Correspondent.

1. Aizenberg, J., Black, A. J. & Whitesides, G. M. *Nature* **398**, 495–498 (1999).
2. Trau, M., Saville, D. A. & Aksay, I. A. *Science* **272**, 706–709 (1996).
3. Hayward, R. C., Saville, D. A. & Aksay, I. A. *Nature* **404**, 56–59 (2000).
4. Belcher, A. M. *et al.* *Nature* **381**, 56–58 (1996).
5. Smith, B. L. *et al.* *Nature* **399**, 761–763 (1999).
6. Whaley, S. R., English, D. S., Hu, E. L., Barbara, P. F. & Belcher, A. M. *Nature* **405**, 665–668 (2000).
7. Fritz, J. *et al.* *Science* **288**, 316–318 (2000).
8. Carlson, R. H. *et al.* *Phys. Rev. Lett.* **79**, 2149–2152 (1997).

'Underachieving' centre has not only struck gold but made good use of it

Sir — In your News feature "The missionary from Munich" (*Nature* **405**, 10; 2000), Rudi Balling, the incoming director of the German Research Centre for Biotechnology (GBF), calls the centre a "gold mine". Yet you, on the other hand, unkindly refer to it as an "underachiever". Presumably this is intended to mean that the gold mine has not yet been properly exploited.

This is not the case. The the GBF mine has not only produced gold bars, but has turned these into numerous pieces of jewellery and sold them to jewellers. Over many years, scientific contributions from this institute have been published in excellent journals, not least in *Nature* (see Hattori *et al.* *Nature* **405**, 311; 2000).

The high standard of GBF's research and development is also reflected by an internal evaluation of the German Helmholtz centres, in which the GBF was one of the top-scoring institutes by all the criteria used. In addition to this, the December 99/January 00 issue of *Biotech International* compared a long list of biotech companies (such as Bayer and the Roche Group) and research institutions (such as the Max Planck Society) outside the United States: the GBF was ranked twenty-fourth. If normalized according to number of personnel, the GBF, which has only 600 employees, would appear in the top ten. Hardly the performance of an underachiever.

The GBF does have one major drawback. Its full name — Gesellschaft für biotechnologische Forschung mit beschränkter Haftung — doesn't exactly roll off German tongues, let alone those that don't speak the language. How much easier life would be if the GBF were, for example, renamed the Eigen Centre (after Nobel laureate Manfred Eigen, one of its main founders). But although a name change might raise the institute's public profile, high visibility should not be confused with high quality.

Helmut Blöcker

Genome Analysis, GBF, Mascheroder Weg 1,
D-38124 Braunschweig, Germany

Jumping the gun on mouse gene expression

Sir — I am writing to correct a statement in the informative News Feature "A silence that speaks volumes" (*Nature* **404**, 804; 2000). The work from my laboratory, published in the paper by Florence Wianny and myself, has shown that dsRNA

interference can be effective in both the mouse oocyte and early embryo (see Wianny, F. & Zernicka-Goetz, M. *Nature Cell Biol.* **2**, 70–75; 2000). We have successfully used the technique to prevent expression of the *c-mos* gene during oocyte maturation, and of the E-cadherin gene to interfere with formation of the blastocyst when delivered to the one-cell-stage zygote.

We are now planning to attempt to interfere at later stages with the expression of genes that help determine polarity during mouse development. However, although we believe that RNAi offers significant potential for such applications, this idea still remains to be demonstrated.

Magdalena Zernicka-Goetz

Wellcome/CRC Institute, Tennis Court Road,
Cambridge CB2 1QR, UK

Setbacks don't dampen the energy of US physics

Sir — I read with great interest Colin Macilwain's News report "Budget crisis forces hard choices on US high-energy physics" (*Nature* **404**, 909; 2000). The report goes a long way in describing the general situation of high-energy physics in the United States. Readers may, however, get the wrong impression, both as to its present intellectual health (despite the severe blow of losing the Superconducting Super Collider in 1993), and as to the efforts being made to get back on track — as the cartoon accompanying the article so dramatically displays.

High-energy physics here is very vigorous, with activity continuing at the Tevatron (Fermilab), activities under way in neutrino physics, use of the new B-Factory and the development of very large detectors to be used at the large hadron collider in CERN.

New facilities being contemplated include an electron-positron linear collider, a very large hadron collider and a neutrino factory. These facilities address very different aspects of high-energy physics. They are not in competition; they all should be built and — with sufficient international cooperation and an adequately long time frame — they will all be built.

The community, working with international colleagues, is developing a suitable plan for high-energy physics. Arriving at a well-thought-out and widely accepted programme for the future is expected to take a few years; a significant step will be undertaken at a retreat next summer.

With regard to budget realities, a neutrino factory may be attractive as a first step: it fits on several existing sites, it would be of worldwide interest, the accelerator cost can be balanced with detector costs and it can be staged. It would also go a long way

towards demonstrating the technologies needed for a possible future muon collider. As we are still improving the design, it is premature to say if it will ultimately cost more or less than the \$1 billion stated in your article. However, our first-draft design could be built in a few increments, each one costing under \$1 billion, and each addressing new, and most interesting, physics.

Andrew M. Sessler

The Neutrino Factory and Muon Collider
Collaboration, Lawrence Berkeley National
Laboratory, University of California, Berkeley,
California 94720, USA

How the GM industry writes its own rules

Sir — Colin Macilwain's News story about the US Food and Drug Administration 'reforms' in the supervision of genetically modified food (*Nature* **405**, 108; 2000) tells only part of the story.

Not only did the announcement "not include any mandatory requirement for the labelling of GM foods" — it did not include any requirement for pre-marketing scientific assessment by the government either.

As a result, we have no real supervision of the industry here. All the FDA did was to make mandatory the 'consultations' which had previously been optional (not a tough demand, since in every case so far, the companies had met FDA officials voluntarily anyway).

The information that is to be placed on the FDA's website will be highly censored, because the United States enforces strict rules protecting proprietary business information. This outweighs any notion of the public's right to know what we are putting in our bodies.

The agency's plan to develop guidelines for voluntary labelling is nice but unnecessary, because our First Amendment guarantees that, as long as a statement is truthful, we don't really need government permission to say it.

This is an exercise in how to tread water and make it appear that you are swimming.

Philip L. Bereano

Council for Responsible Genetics, 5 Upland
Road, Cambridge, Massachusetts 02140, USA

W. D. Hamilton memorial

Sir — A non-religious memorial event for W. D. Hamilton (see *Nature* **404**, 828; 2000) will be held in the chapel of New College, Holywell Street, Oxford, on 1 July 2000 at 2.30 p.m. All are welcome to attend.

Richard Dawkins

New College, Oxford OX1 3BN, UK

When Galileo turned to verse

The English translation of an obscure, 400-year-old Italian curiosity.

Against the Donning of the Gown: *Contro Il Portar La Toga*

by Galileo Galilei, translated by
Giovanni F. Bignami
Moon Books: 2000. 85 pp.

Dava Sobel

Rightly hailed as the father of modern physics and a mighty expositor of Italian prose, Galileo Galilei (1564–1642) also dabbled in poetry. Like other cultured gentlemen of his acquaintance, Galileo composed verses in *terza rima*, the 11-syllable-line style of Dante's *The Divine Comedy*, and wrote sonnets, too. But although all of Galileo's published books on natural philosophy appeared in English translations before the end of the seventeenth century, his poetry has remained an obscure Italian curiosity.

Now, at last, Giovanni Bignami, director of science at the Italian Space Agency, has translated two of Galileo's best poems into English. The result is a book as distinguished for its physical beauty as for the quality of its text (see <http://www.galileounaluna.com>). It looks to be an artefact of Galileo's own time, printed in a limited edition of only 2,000 copies, on "natural cotton fiber paper with neutral pH for conservation beyond time and with uncut edges", according to a typographer's note at the back. It is bound by hand in thick leather covers the colour of vanilla ice-cream, with a facsimile of Galileo's signature impressed on the face, not to mention a woven bookmark sewn into the binding. Even the imprimatur on the title page — the coat of arms of the Lyncean Academy — pulls a thread through history, for in 1611 Galileo joined this early scientific society, which published two of his astronomy books, and today it counts Bignami among its members.

"This is not to be taken seriously," Bignami states in his opening "Caveat" to the poetry. Indeed, Galileo might well have said the same. The first poem, probably written in 1590, presents the youthful Galileo at his most playful, before his father's death in 1591 left him financially responsible for the support of his extended family on an upstart professor's paltry salary, and before the Church Edict of 1616 constrained his broad view of the Universe. This title poem, *Against the Donning of the Gown*, pokes 301 lines of fun at the pompous practice of the University of Pisa, where Galileo began teaching in 1589, of forcing professors to wear academic gowns whenever they appeared in public. Galileo paid frequent fines for breaking this rule. In the poem he claims that people could better appreciate one another's true virtues if everyone were to go naked, and that a man's



Galileo the poet: he wrote that people could better appreciate one another's virtues if everyone were to go naked.

dress tells no more about his true capabilities than a fancy flask discloses of the wine inside it. Despite the light tone and topic, the poem epitomizes Galileo's gusto for a good argument.

Bignami has managed to retain the braided rhyme scheme of *terza rima* (ABA, BCB, CDC, and so on) in his translation. This is a remarkable feat, considering how Italian — a musical language of only seven vowel sounds that rhyme almost effortlessly (even accidentally in ordinary speech) — compares with English, which has 52 vowel sounds. The rhyming does not make the poem feel any more stretched or far-fetched in English than it sounds in Italian, however. Its original mix of highfalutin declamatory airs and low humour is well preserved.

Although *Against the Donning of the Gown* says nothing about science *per se*, the second poem in this volume addresses a true scientific conundrum that eluded Galileo all his life and fuelled some of his nastiest contests with Jesuit contemporaries. This sonnet, "Enigma", poses a riddle to the reader: "Monster I am," the verse begins, "stranger in shape and form/Th[a]n the harpy, the siren or the ghoul".

What is this wretched, wraith-like creature who dwells in the dark, pursued by swarms of hunters? Galileo never published the puzzle's answer, but Bignami shows his own (no doubt correct) solution in an image of a comet drawn by Donata Almicci.

Seventeenth-century philosophers passionately debated the nature of comets in that heady era of early telescopic observa-

tions. Although illness prevented Galileo from viewing the comets that appeared soon after he had perfected the telescope, he denounced all comets as optical illusions conjured up in the Earth's atmosphere. "I watch my limbs disjoin and lose the fight," the closing lines of "Enigma" lament, "As life and name and soul give up I must."

Dava Sobel can be contacted c/o Carlisle & Co., 24 East 64th Street, New York, New York 10021, USA.

Aping human societies

Hierarchy in the Forest: The Evolution of Egalitarian Behavior

by Christopher Boehm
Harvard University Press: 1999. 320 pp.
\$39.95, £24.95

Adrienne Zihlman

Are we by nature hierarchical or egalitarian? In one attempt to answer this question, Christopher Boehm seeks to explore the origins of our egalitarianism. Egalitarian societies are those that act collectively as a moral community to control social and political life. These are the kinds of societies enjoyed by democratic countries today, and before that by foragers living in small bands, by tribes of pastoral nomads and even by some chiefdoms.



Ape-shape: the roots of modern human social behaviour might be found among the Congo chimps.

Boehm ranks human and prehuman groups on a scale ranging from despotism to egalitarianism. His thesis is that human societies evolved from a despotic ape ancestor in a society in which alpha individuals exerted dominance over the others, and subordinates reacted with hostility. From this aggressive beginning, a mix of political arrangements evolved. Boehm's task is to analyse how a chain of evolutionary events has shaped human nature.

The analysis is based on ethnographic work by Boehm and other anthropologists and on Boehm's observations of chimpanzee behaviour at Gombe, Tanzania. In contrast to the hierarchical apes, dominated by their alpha males, the nomadic human foraging bands and sedentary tribal groups use a variety of social controls to prevent such domination by bullies or big egos: disapproval, ridicule, ostracism and, when all else fails, assassination. The quality most valued in a leader is generosity; maintaining successful leadership depends upon listening, consulting, seeking consensus and keeping a low profile. The political structure of such societies is a 'reverse-dominance hierarchy', in which despotic human nature is turned to radically new political use.

The arguments in the first half of this well-written volume are original, persuasive and richly supported by examples from Boehm's personal experience, observations and mastery of the ethnographic literature. But when, in his quest for the origins of modern human behaviour, Boehm ventures into the forest to examine the three living species of African great apes — *Gorilla gorilla*, *Pan troglodytes* and *Pan paniscus* — he is on much soggier ground.

In this twilight zone, we find ourselves transported back several decades to the many versions of the hunting hypothesis that link hunting and eating meat to the origin of

sharing, a sexual division of labour, a big brain, language, and politics. In support of this Big Picture of how we evolved, Boehm focuses on the despotism of the African apes — from the mountain gorilla male guarding his harem, to the aggressiveness, occasionally lethal, of Gombe chimpanzees — which he believes serve as the model for the Common Ancestor. Boehm dismisses the other Congo Basin population of chimpanzees, *P. paniscus*, which are less hierarchical, less aggressive, less male dominated and more affiliative.

But neither mountain gorillas nor Gombe chimpanzees are broadly representative of their species. Mountain gorillas lie at the eastern and altitudinal extremes of their range, although they have been studied more intensively than the more numerous western lowland gorillas. Gombe chimpanzees cover the easternmost range of this widely distributed species, in an environment under stress from habitat encroachment by humans — a plausible factor in their observed aggressiveness.

If Boehm wanted a parsimonious ape model for his egalitarian bands of human hunter-gatherers, he could have started with *P. paniscus* and saved himself the trouble of explaining how humans evolved from an extreme *P. troglodytes* paradigm. *Pan paniscus* shares food, does not have lethal intergroup encounters and is similar to humans in having small canine teeth and little female-male size difference. This anatomy tends to refute Boehm's argument that reduction in canine size is explained by the origin of weapons for hunting, which rendered big canines redundant.

It would seem more logical to treat the two extant chimpanzee species as spanning a broad range of 'chimpanzee politics' — from 'might is right' at Gombe by Lake Tanganyika to 'make love not war' at Wamba in the Congo Basin. As both species are equally

closely related to *Homo sapiens*, there is no *a priori* reason to choose one rather than the other as the progenitor of 'human nature', whatever that may be.

Can we really trace the origin of human behaviour, with its many complex manifestations, through recorded time to *Homo erectus* of half a million years ago, the australopithecines and the five-million-year-old ape ancestor, of which virtually nothing is known? I'm not convinced that we can, and Boehm's speculations along these lines do little to persuade me.

But in seeking the sources of egalitarian societies, Boehm has taken up an important question. I was fascinated by his life among the Navaho and the Montenegrins, and his candid account of his own mistakes in trying to illustrate the sophisticated controls at work in these societies. Such societies and the many others he cites vary in the details of their operation, just as the ape and other primate societies that have been studied vary from one to another. Ambivalence, compromise, the eternal conflict between the individual urge for dominance and the social desire for equality, are worth investigation. But the forest primeval, populated by hypothetical ape ancestors possessing brains one-third the size of our own, may not be the best place to search for the answers. ■

Adrienne Zihlman is in the Department of Anthropology, University of California at Santa Cruz, California 95064, USA.

A helping hand on elementary matters

Lucifer's Legacy: The Meaning of Asymmetry

by Frank Close

Oxford University Press: 2000. 268 pp.
£16.99, \$27.50

P. W. Anderson

In this ambitious book, Frank Close attempts no less a task than to bring the general reader all the way from zero (assumed) knowledge of modern physics to an understanding of the present status and of some crucial questions in the physics of elementary particles. Close's writing is beguiling, mingling personal and historical anecdote with carefully measured doses of exposition in such a way as to guide the reader painlessly into rather deep intellectual waters.

Close chooses to base this enterprise on the idea of symmetry (and the loss thereof), which is an excellent choice. The theories of modern physics can be thought of, to a great extent, as 'applied symmetry'; they are permeated with, and in many ways even based upon, arguments using the symmetries of space and time. Yet, as Close emphasizes, the

very existence of the observable Universe depends upon the spontaneous manifestation of asymmetry, the phenomenon physicists call 'broken symmetry'. Close focuses, at first, on two discrete broken symmetries: the universal 'handedness' of life — its left-right asymmetry — and the cosmic preference for matter over antimatter. In the final chapters, he goes on to expound the broken continuous symmetry that is behind the unification of weak and electromagnetic interactions.

It is hard to judge how accessible the layman will find tutorial explication of the sort contained in this book. For instance, it's not clear to me why Stephen Hawking's writing seems to be comprehensible, or at least acceptable, to the layman; I would judge Close to be much more accessible. So, to my limited sense of these things, this should be a book that the interested layman will enjoy.

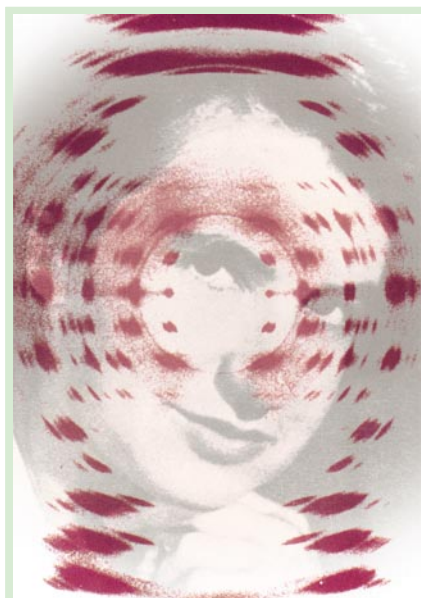
But here's the rub. As I read on, I became increasingly uncomfortable, and particularly so in the final chapters. The first rule of popular writing about science should parallel the well-known law of medicine: 'First, do no harm', which in this context translates into, 'First, get the story straight'.

There are three noteworthy failures to do this. Two major ones may derive from the praiseworthy motivation of making the presentation easier or more dramatic, but as the 'first rule' points out, the baby should not go out with the bath water.

In a book focusing, as this does, on symmetry, it seems misleading not to explain the fundamental principle that all interaction follows from symmetry: the gauge principle of London and Weyl, modelled on and foreshadowed by Einstein's derivation of gravity from general relativity (Einstein seems to be at the root of everything). The beautiful idea that every continuous symmetry implies a conservation law, and an accompanying interaction between the conserved charges, determines the structure of all of the interactions of physics. It is not appropriate to try to approach advanced topics such as electroweak unification and supersymmetry without this foundation block.

Second, there is little resemblance between the real story of how broken symmetry entered particle physics — and of the 'Higgs phenomenon' as the generator of mass — and the rather dramatic little myth created by Close around these events. One supposes that the classic stories of discovery, some of which Close retells elsewhere in the book, may traditionally be somewhat re-touched to spice up the exposition; but that is no excuse for making up a new myth out of whole cloth.

The idea of broken symmetry as a way of generating mass is universally credited to Yoichiro Nambu, who was inspired in 1959–60 by the then new BCS theory of superconductivity. The Higgs phenomenon was, in fact, discovered in that theory by me



A double-take on nature's helix

The striking symmetry of the X-ray crystallograph of DNA prepared by Rosalind Franklin was one of the last clues that led Watson and Crick to deduce the left-handed double-helical structure of DNA. From *Nature's Connections: An Exploration of Natural History* by Nicola McGirr (Natural History Museum, £12.95).

and applied to particle physics in 1963, a year before Higgs' "great inspiration". The major portion of the credit has correctly gone to Steve Weinberg, Abdus Salam, Sheldon Glashow and John Ward (recently deceased), who, around 1967, put all these ideas together in the right way. (A revelatory hint that Close has not understood the mechanics of the Goldstone–Higgs process is the use of the false analogy to ferromagnetism; unlike the ferromagnetic moment, the Higgs field is not a constant of the motion!) Higgs, whom Close measures up for a Nobel, was a rather minor player, whose contribution was parallel to that of Robert Brout and François Englert.

A precedent for Close's emphasis on the Higgs particle, if not on Peter Higgs himself, is the older attempt to cover some of the same material, Leon Lederman's *The God Particle* (1990), in which the eponymous particle is indeed the Higgs. Close will leave the reader much less confused than will Lederman, and is wonderfully frugal with irrelevant detail; but the account of the physics has similar weaknesses.

Finally, while the attention of a few biologists has been caught by the physicists' natural conjecture that the handedness of life is somehow affected by that of the weak interactions, I sense that most thinkers on evolution feel that life would be much harder

to understand if it were not universally chiral.

I personally think it a great pity that Close did not tell the story of broken symmetry as it really happened, because it is a heartening story of one of those rare periods when the fragmentation of theoretical physics into condensed-matter, nuclear and particle branches was temporarily healed and we were all consciously working together in exploring the many quantum consequences of the idea of broken symmetry.

These consequences are as far-reaching and as intellectually stimulating in condensed matter as in particle physics, and in not discussing them Close reveals a definite bias in favour of the 'clean machines' of particle physics as opposed to the mundane complications of condensed matter. Not surprisingly, the book ends with a paean in praise of the magnificent (and magnificently expensive) particle physics machinery at CERN, in happy anticipation of a continued flow of exciting results. One must, of course, share his hope, if not necessarily his optimism. But I regret very deeply the missed opportunity to demonstrate to laymen the unity of the physics enterprise, and Close's treatment of the rest of physics as a poor relation of particle physics, of simply historical value.

P. W. Anderson is in the Department of Physics, Princeton University, PO Box 708, Princeton, New Jersey 08544-0708, USA.

Portrait of an élite world

Cavendish: The Experimental Life by Christa Jungnickel and Russell McCormmach
Bucknell University Press: 1999. 814 pp.
£28, \$34.50

John Gascoigne

The name 'Cavendish' immediately calls to mind the Cambridge laboratory that has produced more Nobel prizewinners than any other scientific institution. There is a link between the laboratory and the late-eighteenth-century Henry Cavendish who is the subject of this definitive biography, but it is an indirect one. When, in 1874, the Duke of Devonshire displayed his aristocratic munificence in founding the laboratory, he chose to commemorate the scientific achievements of his distant relative.

It was an appropriate memorial, as family tradition played an important part in shaping Henry Cavendish's scientific endeavours. The biography brings this aspect out well by taking as its subject not only Henry but also his little-known father, Charles. As the grandson of two dukes

book reviews

(Devonshire and Kent), young Henry had no need to trouble himself with earning a living, and as he grew older he grew richer, thanks to family legacies, a frugal lifestyle and secure investments. He became, as the French physicist Jean Biot put it, “the wisest of the rich and the richest of the wise”.

But he came from an aristocratic élite whose long survival owed much to the fact that they generally understood that privilege had to be balanced by some form of service. The most common form taken by such civic duty was to enter politics, as Henry's father had done. But such a course was closed to Henry by a chronic shyness that led to him literally fleeing if he was button-holed at a social gathering and by a morbid fear of female company.

So, with paternal blessing, Henry turned to science and, in particular, to experimental science of the kind cultivated by his father. He also continued his father's tradition of service to the Royal Society and other learned institutions such as the British Museum. Indeed, it is one of the paradoxes of Cavendish's career that such an extremely introverted character should have been a good committee man. Not only did he serve the Royal Society in various capacities, but he was also a manager of the Royal Institution and a member of the Royal Society of Arts and the Society of Antiquaries. In an age when the characteristic institution of British learned society was the club, the seemingly unclubbable Cavendish was a ubiquitous presence — although one that might rapidly disappear if he was singled out for attention.

Cavendish's chronic shyness was evident, too, in his marked reluctance to publish unless he was totally satisfied with the results — a rare event. The consequence was that, over a lifetime devoted to experimentation (since the bachelor had few domestic distractions), he published fewer than 20 papers. The world, then, saw only some outworks of a majestic edifice that covered most of science as it was then understood. What gave his work unity was his devotion to continuing the Newtonian enterprise of attempting to understand the world in terms of attractive and repulsive forces. It underlay his pioneering work in electricity, which demonstrated both the insights and limitations of the Newtonian concept of force when applied to electrical activity. Although he did publish two papers in this area, largely based on the analogy between the fluid-like behaviour of air and electricity, much of his work, such as his laws of electrical attraction and repulsion, were virtually stillborn as they were never published.

The Newtonian world-view also informed his elegant (and published) 1798 experiment using lead weights on a torsion bar to estimate the strength of the Earth's gravitational attraction which, in turn,



Tools of his trade: lenses used by Cavendish, who viewed science as a form of escape.

enabled Cavendish to calculate the Earth's density and weight.

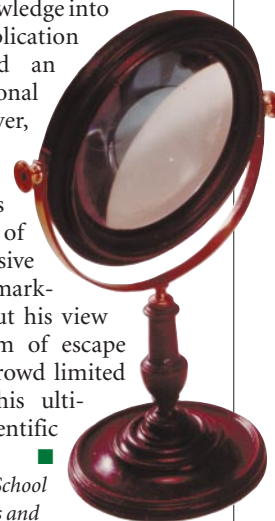
Cavendish's work was also informed by the view that, to some degree, all natural phenomena could be explained in terms of the behaviour of fluids. This was evident both in his electrical experiments and in his pioneering work on what he called ‘factitious airs’, that is, forms of air that can be released from other bodies — of these, the two most notable were ‘inflammable air’ (hydrogen) and ‘fixed air’ (carbon dioxide). The very fact that he distinguished between such forms of air was an indication of the withering of the traditional view that the four elements earth, air, fire and water were the irreducible building-blocks of the world. This ancient understanding of the elements was to be further undermined by Cavendish's experiments in 1784 demonstrating that water was a compound of two airs, ‘inflammable’ and ‘common’. Fortunately, he was sufficiently satisfied with his conclusions to publish them and they were instrumental in prompting

Antoine Lavoisier to develop a totally new understanding of the concept of an element.

In this lengthy work, the authors clearly and systematically describe Cavendish's scientific achievements as well as providing a portrait of the élite world in which he and his father developed their scientific interests. This new edition also publishes Cavendish's surviving scientific correspondence.

After such Herculean labours, it is natural enough for the authors to place their subject on a very high pedestal. Their view that Cavendish “was the preeminent mathematical and experimental scientist in Britain in the century and a half between Newton and Thompson and Maxwell” may be defended, although others might yield such laurels to Thomas Young or Michael Faraday. More contentious is their claim that “Cavendish is one of the greatest scientists ever, as he is one of the most unusual personalities of science”. That claim is perhaps based on an assessment of the whole corpus of his work, with insufficient allowance for the fact that scientific eminence can generally accrue only from published work. For the aristocratic Henry Cavendish, spared the financial pressures to turn knowledge into career capital, publication may have appeared an optional and occasional interruption. However, science is, by its nature, a social activity which depends on the sharing of knowledge. The reclusive Cavendish made a remarkable contribution, but his view of science as a form of escape from the madding crowd limited his influence and his ultimate place in the scientific pantheon.

John Gascoigne is in the School of History, Faculty of Arts and Social Sciences, University of New South Wales, Sydney 2052, Australia.



New in paperback

The River: A Journey Back to the Source of HIV and AIDS

by Edward Hooper
Penguin, £10.99

“*The River* is, in many ways, superb. It is scholarly, thoroughly researched, well (if densely) written and deserves, indeed demands to be taken seriously ... His description of the early days of the African and Western AIDS epidemics is marvellous, but it is his support for the OPV–HIV hypothesis that will attract most attention ... My biggest concern over this book is that it could reinforce public distrust of science and scientists. It is a dangerous policy to hammer

science for unproven — and probably unprovable — events.” John P. Moore, *Nature* **401**, 325–326 (1999)

Evolutionary Wars: A Three-Billion-Year Arms Race

by Charles Kingsley Levy
W. H. Freeman, £10.95, \$16.95

Citizen Science: A Study of People, Expertise and Sustainable Development

by Alan Irwin
Routledge, £16.99, \$24.99

An unholy alliance

The Nazis showed that 'politically responsible' science risks losing its soul.

Ute Deichmann

In 1947, Max Delbrück, the German-born physicist-turned-geneticist then resident in the United States, was preparing to visit Germany. Hermann Muller, a US geneticist, asked him to find out which German geneticists had never voluntarily helped the Nazi regime. This information, he suggested, "would be very helpful to our Committee on Aid to Geneticists Abroad, because many of the members ... do not want to have their money used to help people who had taken part in the prostitution of science".

Seven years earlier, Sir Richard Gregory, former editor of *Nature*, pointed out the dangers of basing science on principles other than scientific ones: "To make race, political convictions, or religious faith, barriers to the pursuit of natural knowledge, means that science in Nazi Germany loses its soul for the purpose of gaining the world." Muller and Gregory believed that science is so pure that its ideological and practical support of Nazi race policy could be called a prostitution, and is so universal and independent of national or ethnic affiliations that it would lose its soul if it excluded people of supposedly different races.

More recently, disillusionment about the contributions of eminent scientists to the Nazi regime has led many to question the notion of a pure, universal, basic science, and even to reject it as a myth. Instead, science has been redefined as a socially organized, political enterprise with a high potential for creating power. Proponents of this view argue that science must be politically responsible, directed towards socially acceptable goals, and assessed according to its long-range consequences.

But the call for politically responsible science, and hence more power for scientists, does not guarantee an ethical stance. Environmentalists' attempts in the 1980s to create a 'political ecology' as the 'guiding science of post-modernism' is a case in point. The intellectual origins of their criticisms of 'causal reductionist' science lie in the 1920s, when German ecologists, among them Karl Friederichs, proclaimed ecology as a path to "a view of the world, in which everything is related to everything else, everything directly or indirectly affects everything else". Friederichs became a leading Nazi ecologist, and he and his colleagues created and spread the view of biology as an eminent political sci-



Loaded questions: a German physician conducts 'ethnological' research on a gypsy woman. Left: Adolf Windaus, who resisted calls to "serve the general good".



ence aimed at serving "the benefit of the people [Volk]" and of ecology as the "doctrine of blood and soil".

Eugenics, or race hygiene, is another case of scientists claiming to act in a politically responsible manner. To avert long-range threats to the gene pool they demanded compulsory sterilization of 'genetically unfit' people. These attempts to create a politically responsible biology ended disastrously. If we criticize reductionist science for having contributed to the technical and military power of the Nazis, we have to acknowledge that 'politically responsible' biologists provided for their ideological and political power.

In contrast to the rhetoric of his fellow German scientists, Adolf Windaus, the 1928 Nobel laureate for chemistry — and one of the few who refused to compromise or work with the Nazis — frankly admitted that his science had no ethical agenda. In 1948, a colleague argued that "it is the final task of every science to serve the general good".

"I have always felt differently," replied Windaus. "My motive has been scientific curiosity. I wanted to know what is the structure of a substance... I have never thought that by this I might serve the general good. ... I was induced to remind myself of all this, because recently a Swiss philosopher strongly criticised scientific curiosity, the quest for knowledge, for the sake of knowledge, as value-free science. He even attributed the atrocities of the experiments on defenceless

humans to such curiosity. I believe a scientist is self-evidently bound to ethical principles, as is every man, but his quest for knowledge has, at first, nothing to do with morality."

Windhaus was right to point to a scientific level outside politics, ethics and applications. It is not the quest for knowledge that was responsible for the atrocities, but, as Windhaus argued, the fact that scientists did not pay due regard to normal ethical principles. Nazi 'moral standards' were not imposed on scientists. On the contrary, for whatever reason — opportunism, conviction, promotion or power — scientists lent their support to ranking human beings as valuable, inferior or worthless, hence providing the ideological basis of the Nazi state.

Otmar von Verschuer, for example, the director of the Kaiser Wilhelm Institute for Anthropology, collaborated with Josef Mengele in Auschwitz. His acceptance of organs and blood from deliberately infected concentration-camp inmates, stands, according to the German geneticist Benno Müller-Hill, as the most infamous crime in which geneticists have participated. These researchers clearly transgressed the limits of science.

The example of Nazi Germany shows that 'politically responsible' science endowed with power can have disastrous consequences for innocent people and for science itself. The call for politically responsible science, frequently heard today, cannot solve the problem of how scientists can prevent science from serving immoral, inhuman ends. ■

Ute Deichmann is a historian of science at the Institut für Genetik, Weyertal 121, D 50931 Köln, Germany.

Another green world

There's more to industrial waste than chimneys and slag heaps.

Henry Wessells

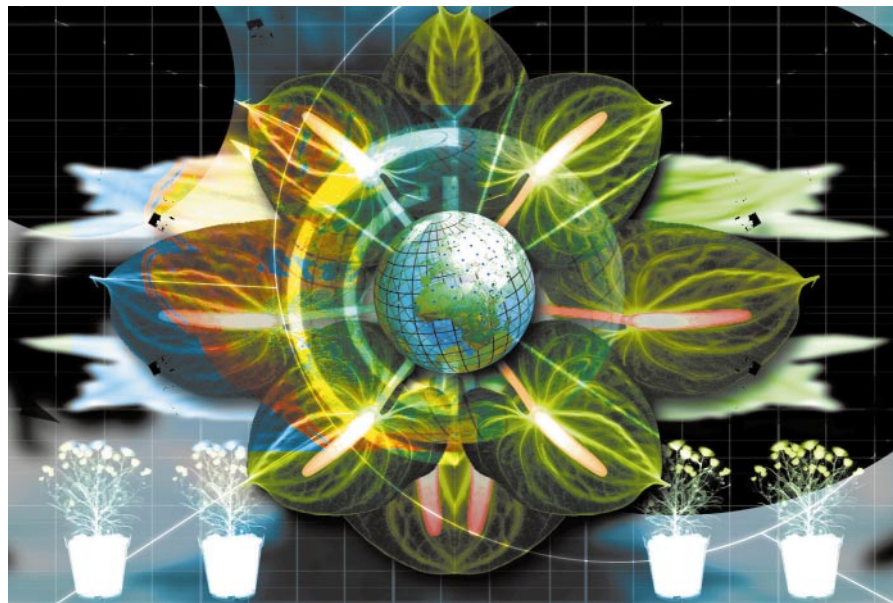
Peter Schmidt should, perhaps, have known better. He was accustomed to taking risks: as a graduate student, he had inserted genes from a potent strain of *Cannabis sativa* into a kudzu vine (*Pueraria thunbergiana*) with spectacular results. For several years, he and his brother Gustav had grown the transgenic plants in artfully derelict warehouses in North Philadelphia. His brother handled distribution; their profits were laundered in secret accounts in a bank on Great Andaman. When one of the warehouses collapsed in a heavy June thunderstorm, the hybrid proved itself capable of crossing with natural varieties of kudzu growing in Wissahickon Park. The rapidly growing vines bearing giant hemp leaves (and significant levels of tetrahydrocannabinol) soon acclimated to Pennsylvania and began to spread. By that time, Schmidt no longer needed to rely on clandestine funding.

His published work was devoted to the theory and practice of increasing heavy-metal uptake by genetically altered plants. His PhD thesis, *Genetic Enhancement of Chromium and Cadmium Uptake in Phragmites sp. and the Rehabilitation of the Parabiosphere* (University of Arizona, 2004) served as the business plan for the launch of Oxygenerator, Inc.

The company's first project, decontamination of marshes in the Meadowlands of northern New Jersey, was only a mixed success, despite early, under-budget completion. The labourers engaged to harvest and process the biomass — refugee Iraqi Marsh Arabs — began to manifest dermatological and teratogenic effects. Schmidt maintained that these were unrelated to the work and should properly be viewed as consequences of the Gulf War and earlier Iraqi repression. Renewed protests outside his Arizona headquarters had turned violent on several occasions.

Undaunted, Schmidt pursued an even more ambitious aim: to develop a multi-purpose biomass producer adaptable both to the reclamation of desert lands in Pakistan and urban wastelands in Camden, Warsaw and Mexico City. Almost on the eve of the project's implementation, a 'temporary readjustment' of global stock markets and the usual bureaucratic delays had created a short-term funding crunch.

That afternoon, as he sat outside his office with Bill Grey, on the terrace overlooking the company greenhouses and the mountains



west of Tucson, he saw a situation that demanded an unconventional solution.

"Bill, your company has been our sole certified waste-disposal contractor since day one, and I really value your services. We couldn't navigate the federal and EU regulations without you. In six months, we're going to be looking at a major expansion in our testing volume, when we get formal approval for full-scale development of a badlands plant to be funded by the UN as part of the Indo-Pakistan Disarmament Treaty. That means, of course, a major increase in the reject material you'll be handling."

Schmidt paused. The only listening devices within range were his own — and he had turned them off personally.

Grey nodded, his craggy face a mask, except for a certain eagerness in his eyes.

"In the interim," Schmidt continued, "there is one aspect of our relationship that needs to be reconsidered. The project doesn't have regulatory approval yet, and so no disposal certification requirements have been drafted. But if I were to initiate even partial testing now, we'd gain much more than simply six months head start. I can't really convey to you how valuable that would be. It is, however, impossible," Schmidt paused again, and his next word lingered, "unless..."

Grey looked down at the table, then quickly raised his eyes. A faint hint of a smile flickered at the right corner of his mouth. "I think our firm can handle your needs in this area. What volume do you anticipate? One truck every week or two for six months? Peter, I started out driving a truck for my father,

and I haven't forgotten how. I know just the stretch of empty land." He chuckled and wrote down a figure on a scrap of paper. "Half up front, the remainder at three months."

Peter Schmidt glanced at the figure. "I already have some unsuccessful manipulations, altered material that has proven unsatisfactory. When can you make the first haul?" He stood up and reached across the table to shake Grey's hand.

A seed sprouting awareness in arid silicate soil, a molecular green language assembling momentary knowledge from stored potential. Language awaiting germination with encoded dreams of eating asphalt and automobiles, concrete and polyvinyl chloride, fibreglass and steel. In the seedsplit greenness of now, there is no meaning to past and future, centre and periphery, subject and object. Sprouting tendrils of a green millennium released from dream into sunshine and possibility. The uncontrolled adaptation of green intelligence assembling molecular data flows from available organic matter. Seeds dispersed to the winds, eaten by birds, crunch and smear onto passing cars in slam-dance methodology of language awaiting circumstance.

Outwardly, nothing has changed yet. It is not unusual for a barren patch of ground to sprout new growth during Tucson's brief wet season.

Henry Wessells is a freelance writer and editor living in Upper Montclair, New Jersey (wessells@aol.com). He is founder of the Avram Davidson Society and co-editor, with Grania Davis, of a forthcoming collection of stories by Avram Davidson (Tor).

Moon, tides and climate

Carl Wunsch

The view that much of the energy of ocean tides is dissipated in deep water, rather than in shallow coastal seas, now finds observational support. Curiously, the results bear upon our understanding of climate change.

The Moon is receding from the Earth at about 4 centimetres a year, as measured¹ by laser reflectors left there by astronauts. What does this motion have to do with the ocean circulation? By Kepler's laws, the recession implies that there is a continuing loss of energy in the Earth–Moon system of 3×10^{12} watts, or 3 terawatts, mostly in the ocean. But where in the ocean does this energy go, and what are its effects? On page 775 of this issue, Egbert and Ray² produce evidence for the seemingly lunatic conclusion³ that dissipation of tidal energy in the deep sea, and the resulting mixing, are controlling features of the overall ocean circulation.

In the conventional picture of the oceans, there is a wind-driven upper circulation that gives rise to massive, near-surface flows such as the Gulf Stream and the Kuroshio and Antarctic Circumpolar currents. Superimposed upon this circulation is one often labelled, unhappily, the 'thermohaline' circulation. This is supposedly driven by surface-ocean density contrasts arising from temperature and salt variations produced by strong atmospheric cooling and wind-induced evaporation. In this process, dense water sinks at high latitudes through convection, driving a 'meridional overturning circulation', which many believe dominates the heat and freshwater budgets of the climate system. (The terminological problem with 'thermohaline' circulation arises because, for example, half the heat transport in the North Pacific Ocean is in the wind-driven upper circulation⁴.)

As formulated in almost all models of the Earth's climate, both theoretical and numerical, the dense water sinking in the meridional overturning circulation at high latitudes then flows, close to the ocean bottom, throughout the world, returning to the surface by a uniform upwelling through the 'interior' ocean. Under the simple assumption that a uniform upwelling of cold water is balanced by a uniform downward mixing of warmer water throughout the water column, a steady state is achieved. Almost all numerical models of the ocean and climate systems represent this process through spatially constant vertical 'eddy'-mixing coefficients, as do the textbook theories.

It has become evident, however, that the actual circulation is much more subtle and interesting. Consideration of the stability and energetics of a fluid being heated and



Figure 1 The Moon was not usually believed to be involved in the general circulation of the oceans. But in Russian literature, Private Kozma Prutkov, regarded as quite dim, usually produces the right answers. When asked which is more important, the Sun or Moon, he replies: "The Moon, of course, because the sun shines only in daytime when it is bright anyhow..." (Drawing by M. Dormer¹⁶.)

cooled at the surface⁵ shows that the resulting motion would be extremely weak — a 'diffusive creep'. Such a fluid system is stable, and in a steady state it cannot produce the vigorous flow we observe in the deep oceans. There cannot be a primarily convectively driven circulation of any significance^{3,6}.

Furthermore, as early as about 1970 it was clear that the picture of uniform vertical mixing was not correct in the upper ocean^{7–9}. Both direct measurements of turbulence and dye-diffusion experiments^{10,11} have shown the weakness of open-ocean mixing all the way down to the sea floor. Instead, measurements confirm an old hypothesis³: that the ocean is mixed primarily at its boundaries, including the mid-ocean ridges that rise from the sea floor, where the local mixing rates are orders of magnitude larger than those in the ocean interior^{10,11}. The reliance of almost all numerical circulation models on uniform interior-ocean mixing calls into question inferences about the physics of the circulation based on them. Only in the past

two years have models that eliminate such mixing¹² finally started to appear.

Surprisingly, it was only recently recognized that the need for an energy source to sustain the vertical mixing (lifting dense water through lighter) has important consequences. The difficulties of driving fluid motions by surface heating and evaporation mean that a mechanical source of energy must control not only the directly wind-driven flows, but also the deep-water components of the meridional overturning circulation. There are only two candidates for such a source: winds and tides.

For over 75 years^{13,14} it was thought that the tides dissipated almost entirely by friction in the shallow seas above the continental shelves. But Munk and I concluded³ that about half of the power required to return the deep waters to the surface was coming from mixing driven primarily by dissipation of tidal energy — principally lunar, but with a minor solar component — in the deep ocean (Fig. 1). Now, by fitting a dynamical model to satellite altimetric measurements of the tides, Egbert and Ray² have produced an observational estimate of 1 terawatt of open-ocean tidal dissipation. Their numbers are not definitive, but they are in agreement with the energy values required by the deep upwelling, and with the total — shallow (about 2 terawatts) plus deep — energy losses implied by the lunar recession.

If the hypothesis of tide- and wind-driven controls on the rates at which the ocean transports heat and fresh water survives further tests, there are several implications. One is that it brings into question the extent to which uniform-mixing models of the ocean circulation could either reproduce the present-day circulation or predict responses to external changes. The hypothesis also suggests that the rate-limiting¹⁵ factor for oceanic heat transport is not primarily the surface density gradient imposed on the ocean; rather, it is the strengths and patterns of the wind, and the distributions of the tides.

What would be the consequences if the hypothesis is correct? One is that the atmospheric wind patterns would have to be known in considering past and future climate change. The other is that changes in tidal distributions and the consequent mixing would need to be understood over geological time. During the Last Glacial Maximum, the sea level was about 130

metres lower than today. This configuration removed much of the present regions of shallow-water energy dissipation and changed the deep-ocean tides, presumably affecting oceanic heat transport. Over longer periods in the past, the entire continental configuration was different, with radically different tidal distributions and mixing. It appears that the tides are, surprisingly, an intricate part of the story of climate change, as is the history of the lunar orbit. ■

Carl Wunsch is in the Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
e-mail: cwunsch@pond.mit.edu

1. Dickey, J. O. *et al.* *Science* **265**, 482–490 (1994).
2. Egbert, G. D. & Ray, R. D. *Nature* **405**, 775–778 (2000).
3. Munk, W. & Wunsch, C. *Deep-Sea Res.* **45**, 1976–2009 (1998).
4. Bryden, H. L., Roemmich, D. H. & Church, J. A. *Deep-Sea Res.* **38**, 297–324 (1991).
5. Sandström, J. W. *Ann. Hydr. Mar. Met.* **6** (1908).
6. Huang, R. X. *J. Phys. Oceanogr.* **29**, 727–746 (1999).
7. Munk, W. H. *Deep-Sea Res.* **13**, 707–730 (1966).
8. Osborn, T. R. & Cox, C. S. *Geophys. Fluid Dyn.* **3**, 321–345 (1972).
9. Gregg, M. C. *J. Geophys. Res.* **92**, 5249–5286 (1987).
10. Ledwell, J. R. *et al.* *Nature* **403**, 179–182 (2000).
11. Polzin, K., Toole, J. M., Ledwell, J. R. & Schmitt, R. W. *Science* **276**, 93–96 (1997).
12. Marotzke, J. *J. Phys. Oceanogr.* **27**, 1713–1728 (1997).
13. Jeffreys, H. *Phil. Trans. R. Soc. Lond. A* **221**, 239–264 (1920).
14. Taylor, G. I. *Phil. Trans. R. Soc. Lond. A* **220**, 1–33 (1919).
15. Marotzke, J. & Scott, J. R. *J. Phys. Oceanogr.* **29**, 2962–2970 (1999).
16. Munk, W. & Wunsch, C. *Oceanography* **10**, 132–134 (1997).

Demography

Greater lifetime expectations

Shiro Horiuchi

For early humans, the average lifespan was around 20 years, as estimated from skeletal remains. Now, in several industrialized countries, it is about 80 years. Much of this increase has happened in the past 150 years. But it was widely expected that as life expectancy became very high and approached the ‘biological limit of human longevity’, the rapid ‘mortality decline’ would slow down and eventually level off.

Such a deceleration has not occurred yet. On page 789 of this issue, Tuljapurkar *et al.*¹ show that during the second half of the twentieth century, age-specific death rates in the G7 industrialized countries — the United States, Canada, Japan, France, Germany, Italy and the United Kingdom — continued to decline at a remarkably constant pace. There was no noticeable sign of a slowing down. This report follows one by Wilmoth², published two years ago, which indicated that the mortality reduction in the United States, measured by the age-standardized death rate, was even faster in the second half of the century than in the first half.

Assuming a further continuation of the stable pace of mortality decline, Tuljapurkar *et al.* forecast that the life expectancy at birth is likely to increase faster than predicted by the governments of the G7 countries (Fig. 1). This implies that the elderly population in the near future will be greater than in the official forecasts. Depending on the general state of health, the larger elderly population could entail higher medical costs and demands for long-term care and other services, and higher pension payments.

The findings give rise to two interrelated questions. Why has mortality decline not started to slow down? And will it continue into the future? Studies in demography, epidemiology and the biology of ageing

	Official forecasts for 2050	Forecasts of Tuljapurkar <i>et al.</i> ¹ for 2050	Gap
United States	80.45	82.91	2.5
Canada	81.67	85.26	3.6
Japan	82.95	90.91	8.0
France	83.50	87.01	3.5
Germany	81.50	83.12	1.6
Italy	82.50	86.26	3.8
United Kingdom	82.50	83.79	1.3

Figure 1 Official medium-variant forecasts of life expectancy in the G7 countries in 2050 compared with the forecasts of Tuljapurkar *et al.*¹. The figures are for life expectancy for the sexes combined. They are taken from Table 3 of the paper on page 792, and those in the ‘gap’ column have been rounded to the nearest decimal point.

and longevity provide clues to the answers. Underlying the steady decrease in mortality level were shifts in the pattern of mortality reduction³. In the second half of the nineteenth century and the first half of the twentieth century, there were large decreases in the number of deaths from infectious and parasitic diseases, and from poor nutrition and disorders associated with pregnancy and childbirth. The reduction was pronounced among infants, children and young adults, but modest among the elderly. This led to a view that the fall in death rates at young and middle ages to low levels would soon exhaust the potential to prolong life expectancy further.

But that view did not — and could not — take account of developments in the second half of the twentieth century. Mortality from degenerative diseases, most notably heart diseases and stroke, started to fall⁴. The reduction was pronounced among the

elderly^{5,6}, and some suspected that it might have been achieved through postponing the deaths of seriously ill people. But in the United States at least^{7,8}, it seems that the health of the elderly greatly improved in the 1980s and 1990s, suggesting that the extended length of life in old age is mainly due to better health rather than prolonged survival in sickness.

Another shift in the pattern of mortality reduction might also have occurred. Despite the marked decrease in deaths from various degenerative diseases, the overall level of cancer mortality remained the same for many years. But, around 1990, a long-awaited decline in total cancer mortality finally started in economically developed countries. Whether that downward trend will continue for long remains to be seen.

These days, the existence of a biological limit to human longevity is considered questionable⁹. Biologists used to think that senescent processes might be programmed into the biological clock of the human body. But they have largely shifted to the view that senescence is mainly due to the body’s imperfect systems of maintenance and repair, which allow the long-term accumulation of unrepaired damage in macromolecules, cells, tissues and organs¹⁰. Progress in ageing research may eventually lead to new medical approaches that lower the rates of damage accumulation¹¹.

Overall, the evidence supports the expectation that scientific, technological and economic developments will lead to more effective control of degenerative diseases and ageing processes, making it possible to sustain the rapid pace of mortality decline. However, this prospect is not unconditional. New threats to health and survival are arising, including the emergence and re-emergence of infectious diseases, increasing pollution, and the proliferation of nuclear, biological and chemical weapons. If we fail to control these hazards, some of the large gain in the life expectancy of the past 150 years may well be lost³.

What about the world outside the G7 nations? The prospect of life expectancy soon exceeding 80 years is limited to these and other countries with highly developed market economies. They are mostly in Western Europe, North America and Eastern Asia. In the rest of the world, the life expectancy is on average still under 65 years¹². In subSaharan Africa, it is under 50 years and may be falling because of the AIDS epidemic, as well as stagnated economic development and political conflicts. Some industrialized nations in Eastern Europe and the former Soviet Union have seen only slow increases, and even occasional reversals, in life expectancy.

It is not surprising that government forecasts of life expectancy and the size of the elderly population in G7 countries are

much more conservative than those of Tuljapurkar *et al.*¹. In the past, national governments, as well as international organizations and academic researchers, have almost invariably underpredicted life expectancy in industrialized market-economy countries. For example, in 1984 the United Nations prepared international population projections based on the assumption that the maximal life expectancy in human populations is 75 years for males and 82.5 years for females. The assumption soon proved wrong: in Japan, for instance, the figures in 1998 were 77.2 years and 84 years respectively. The reason for the underestimates is partly because forecasters regard conservative prospects as less controversial and so 'safe'; and partly because they have often extrapolated past trends in death rates by cause of death or age (or both), missing future transitions in the cause-of-death pattern and age pattern of mortality

reduction. It looks as if the same error is still being made.

Shiro Horiuchi is in the Laboratory of Populations, Rockefeller University, 1230 York Avenue, Box 20, New York, New York 10021-6399, USA.
e-mail: horiush@rockefeller.edu

1. Tuljapurkar, S., Li, N. & Boe, C. *Nature* **405**, 789–792 (2000).
2. Wilmoth, J. R. *Science* **280**, 395–397 (1998).
3. Horiuchi, S. in *Health and Mortality: Issues of Global Concern* 54–71 (United Nations, New York, 1999).
4. Olshansky, S. J. & Ault, A. B. *Milbank Mem. Fund Q.* **64**, 355–391 (1986).
5. Vaupel, J. W. *Phil. Trans. R. Soc. Lond. B* **352**, 1799–1804 (1997).
6. Wilmoth, J. R. & Horiuchi, S. *Demography* **36**, 475–495 (1999).
7. Manton, K., Corder, L. S. & Stallard, E. *Proc. Natl Acad. Sci. USA* **94**, 2593–2598 (1997).
8. Crimmins, E. M., Saito, Y. & Ingegneri, D. *Pop. Dev. Rev.* **23**, 555–572 (1997).
9. Wilmoth, J. R. in *Between Zeus and the Salmon: The Biodemography of Longevity* (eds Wachter, K. E. & Finch, C. E.) 38–64 (National Academy Press, Washington DC, 1997).
10. Kirkwood, T. B. L. *Time of Our Lives: The Science of Human Aging* (Oxford Univ. Press, 1999).
11. Miller, R. A. *J. Am. Geriatr. Soc.* **45**, 1258–1267 (1997).
12. *World Population Prospects: The 1998 Revision Vol. 1 Comprehensive Tables* (United Nations, New York, 1999).

Fluid dynamics

Smart polymer solutions

Jacob Klein

The proverbial water stays off a duck's back because the duck has waxy water-repellent layers on its feathers. Similarly hydrophobic layers on the surface of most leaves ensure that raindrops bounce or roll off them¹. This is sometimes known as the Lotus effect — after the leaves of the Lotus flower, which are particularly hydrophobic — and makes it easier for rain to wash away dust and dirt and keep the foliage healthy². But these same waxy layers can have less desirable consequences: for example, they cause droplets of herbicide and pesticide sprays to bounce off plants³. As a consequence, over 50% of these toxic sprays are wasted, making it difficult to protect crops and meet environmental regulations.

On page 772 of this issue, Bergeron and co-workers⁴ show that by adding low concentrations of long, flexible polymers to water droplets, they can prevent the droplets from bouncing off hydrophobic surfaces. As well as showing how plants could be more effectively sprayed, the work has applications in other areas, such as the efficiency of water-based inks or paints to coat water-repellent surfaces. In this case, polymer additives would eliminate the unwanted side effects associated with the use of noxious organic solvents, which wet such surfaces more readily.

Liquid drops hitting a flat, solid surface expand and flatten, because of their momentum. Most studies to date^{5,6} — driven by the need to coat surfaces with liquids as efficiently as possible — have focused on the behaviour

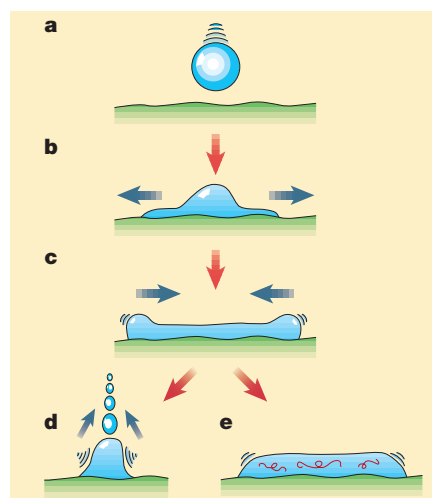


Figure 1 How to save on weedkiller. a–c, The progressive flattening of a water drop hitting a solid surface. c, The drop has reached its maximum radius and begins to retract from the hydrophobic surface. d, Pure water then retracts rapidly enough to be ejected from the surface. e, Bergeron *et al.*⁴ show that adding a low concentration of long, flexible polymers slows down retraction — as a result of increasing elongational viscosity — to the extent that ejection does not occur.

of so-called newtonian fluids⁷, such as water, whose viscosity is essentially unchanged by most rates of deformation or 'shear'. In the case of water hitting a hydrophobic material, the drops retract after flattening to minimize their contact with the surface; this retraction

may be so rapid that part of the drop is ejected and lost from the surface (Fig. 1). Bergeron *et al.*⁴ found that a small concentration of a long, flexible polymer added to the water dramatically slows down the retraction rate of the flattened drop on a hydrophobic surface. This prevents it from reaching 'escape velocity', ensuring that the drop — and anything delivered within it — remains on the surface.

Why does the addition of such small amounts of polymer (typically 0.1 g of polymer per litre of water) have such a large effect? One way that very low concentrations of polymer additives could produce a large surface effect is through adsorption, by which the polymers adhere to the solid surface. Indeed, studies have shown that polymers attached to non-wetting surfaces can stabilize liquid films and prevent 'dewetting'^{8,9}. A simple calculation suggests that the duration of the droplet spreading in the experiments of Bergeron *et al.*⁴ is sufficient for a polymer layer to adsorb onto the surface. But the authors report that the surfaces remain similarly hydrophobic both before and after the polymer solution has made contact with them.

From this, they conclude that adsorption of the polymer — which would reduce the water repellency and would therefore slow down the drop retraction — is negligible and is unlikely to provide the answer. In fact, they suggest a different origin for the slowing down of the retraction phase. They propose that the flexible polymer coils provide a large resistance to being stretched in the rapidly deforming drop as it retracts after spreading. Such resistance results in a drag on the flowing liquid that is known as elongational or extensional viscosity⁷, and it is this, they suggest, that is responsible for the sluggish retraction rate and that stops the drop from bouncing off. The initial flattening is dominated by the drop's momentum, and is much less affected by the elongational viscosity. But the retraction is driven by hydrophobicity, a weaker driving force, which is therefore more affected.

Elongational viscosity in polymer solutions is a non-newtonian effect that comes into play only at deformation rates comparable to or larger than the molecular relaxation rates of the polymer molecules. It can then be orders of magnitude larger than the more familiar shear viscosity, which arises from the drag of the solvent on the undistorted polymer coils and manifests itself at all deformation rates. Elongational viscosity has long been exploited in firefighting, where a tiny concentration of dissolved polymers can dramatically increase the range of the water jet emerging from the hose.

This effect, known as turbulent drag reduction¹⁰, is thought to result mainly from an increase in elongational viscosity, which suppresses the spread of turbulence and so

ensures that the flow within the pipe remains streamlined and rapid (an extra benefit is suppression of the jet break-up as it leaves the nozzle). This is one of the few known practical applications of elongational viscosity, as opposed to the widespread use of polymers to increase shear viscosity, such as the thickening of foodstuffs by starch, a naturally occurring polymer. So it is intriguing, as noted by Bergeron *et al.*, that the same elongational viscosity should act to reduce rather than increase flow, thereby suppressing the rebound of water droplets.

To confirm their theory, Bergeron and co-workers⁴ show that drops of a polymer solution with a high extensional viscosity retract at the same speed as drops of polymer-free water thickened to the same high viscosity by mixing with glycerol. Such evidence is circumstantial but persuasive: the implication is a new and unexpected demonstration of extensional viscosity, and may find applications in other areas where rapidly deforming thin films are involved. An extra bonus of this system, as the authors point out, is that the shear viscosity of their

dilute solutions remains almost indistinguishable from that of pure water, ensuring that little viscous friction is encountered while handling and pumping the liquids. The high elongational viscosity becomes important only when it is needed — during the rapid deformation of the retracting drop. Not an intelligent liquid, perhaps, in view of its simplicity, but certainly a smart one. ■

Jacob Klein is in the Department of Materials and Interfaces, Weizmann Institute of Science, Rehovot 76100, Israel.

e-mail: jacob.klein@weizmann.ac.il

1. Neinhuis, C. & Barthlott, W. *Ann. Bot.* **79**, 667–677 (1997).
2. Barthlott, W. & Neinhuis, C. *Planta* **202**, 1–8 (1997).
3. Wirth, W., Storp, S. & Jacobsen, W. *Pest. Sci.* **33**, 411–422 (1991).
4. Bergeron, V., Bonn, D., Martin, J. Y. & Vovelle, L. *Nature* **405**, 772–775 (2000).
5. Chandra, S. & Avedisian, C. T. *Proc. R. Soc. Lond. A* **432**, 13–41 (1991).
6. Mao, T., Kuhn, D. & Tran, H. *AIChE J.* **43**, 2169–2179 (1997).
7. Bird, R. B., Armstrong, R. C. & Hassager, O. *Dynamics of Polymeric Liquids Vols 1 & 2* (Wiley, New York, 1987).
8. Yerushalmi-Rozen, R., Klein, J. & Fetters, L. J. *Science* **263**, 793–795 (1994).
9. Yerushalmi-Rozen, R. & Klein, J. *Langmuir* **11**, 2806–2814 (1995).
10. Gyr, A. & Bewersdorff, H. W. *Drag Reduction in Turbulent Flows by Additives* (Kluwer, London, 1995).

Neurobiology

Nervous engineering

Melitta Schachner

The ability to repair damaged tissue in the human nervous system has long been a goal of neurobiologists. Writing in *Proceedings of the National Academy of Sciences*, Todd Holmes and colleagues¹ describe a promising biomaterial for this purpose. It consists of a peptide scaffold that can act as a substrate for the attachment of neurons, and allows the growth of nerve fibres and the formation of synapses — the specialized connections between neurons. It is early days as yet, but the idea is that such artificially grown tissue could be transplanted into patients.

Repairing nervous tissue is no easy task. Cell replenishment happens in many other adult mammalian tissues, but very few new neurons are produced in the adult central nervous system. Moreover, the outgrowth of fibres from regenerating neurons into damaged areas is controlled by various molecules that can promote or inhibit the process; and neurons that lack an appropriate substrate cannot regrow and are vulnerable to self-destruction through apoptosis. An ideal transplantable substrate for repairing damaged tissue in the nervous system should support neuronal attachment, fibre outgrowth, and survival and formation of active synapses. Such a substrate should be well tolerated *in vivo*.

Development of the new peptide-scaffold biomaterials¹ started with a serendipitous

observation by Holmes and Shuguang Zhang, another of the authors on the paper. Holmes was testing the neurotoxicity of peptides in neuronal cultures, and Zhang provided him with the so-called EAK16 peptide for the purpose. As its name implies, EAK16 is 16 amino acids long, and it is made up of repeating units of negatively charged

glutamate residues (E in single-letter amino-acid code) and positively charged lysines (K), separated by hydrophobic alanines (A). This arrangement gives EAK16 two distinct polar and non-polar surfaces.

Much to their surprise, Holmes and Zhang observed the formation of macroscopically well-ordered, thin-sheet structures in the neuronal cultures into which EAK16 was introduced; moreover, there was no measurable neurotoxicity in the cultures exposed to the peptide. They went on to find out that the formation of the macroscopic sheet structures from EAK16 depended on millimolar levels of monovalent salts. Scanning electron microscopy of the sheet structures revealed a fibrous assembly of the EAK16 material. The openings between the microscopic fibrils were small enough to exclude cells, but large enough to allow the passage of macromolecules. Holmes, Zhang and colleagues devised a molecular model for the salt-induced formation of the EAK16 material and published their findings in 1993 (ref. 2).

The sequence of EAK16 has some similarity to the RGD (arginine-glycine-aspartate) sequence that is characteristic of some integrin receptors, molecules that are central players in cell adhesion and nerve-fibre outgrowth. On making the EAK16 derivatives RGD16 and RAD16, Holmes and Zhang found that the RAD16 peptide formed stable macroscopic sheets in solutions containing salt at physiological levels. In contrast, no macroscopic materials were formed from the RGD peptides. With the help of Michael DiPersio, they tested the hypothesis that cells would attach and grow on the RAD16 peptide biomaterials in an integrin-dependent fashion. They found, however, that both EAK16 and RAD16 peptide biomaterials robustly supported the attachment and growth of many types of non-neural primary

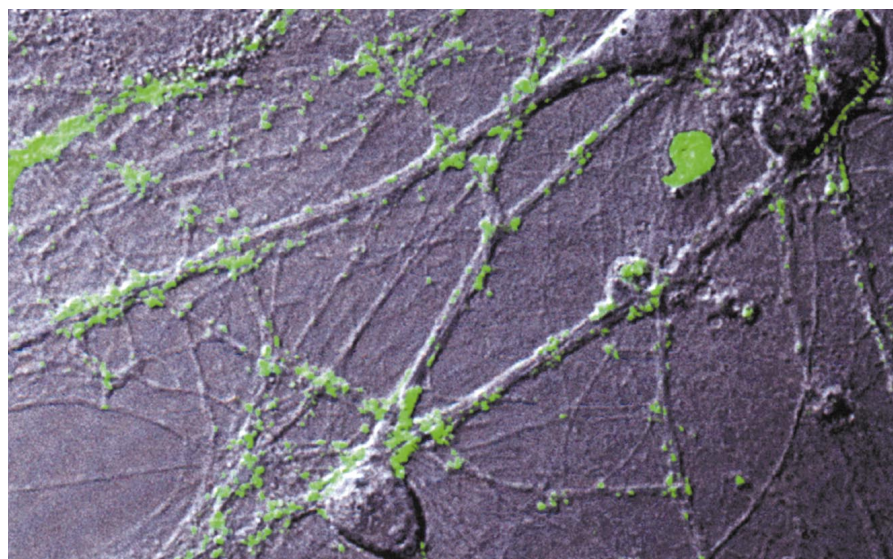


Figure 1 Synaptic activity of neurons grown by Holmes *et al.*¹ on their peptide scaffolds. The neurons concerned come from rat hippocampus; the fluorescent dye is indicative of neurotransmitter release. (Reproduced from ref. 1.)

and transformed cells³. Many of the cell types completely colonized all surfaces of the peptide-based sheet biomaterials, giving the appearance of tissue.

Remarkably, the new paper¹ shows that these peptide biomaterials also support neuronal survival. Robust networks of nerve fibres grew from both primary neurons and neuron-like transformed cells on the RAD16 scaffold, and to a lesser extent on the EAK16 scaffold. Furthermore, the neurons formed functional synapses, as shown by labelling with a fluorescent dye that was taken up into vesicles following neurotransmitter release at synapses (Fig. 1). There were no immunological or inflammatory reactions when the peptides were injected into rat muscle.

Inflammatory reactions in muscle and brain are quite different, however. So these peptide-based biomaterials will have to succeed in tests of how well they are tolerated in brain, spinal cord and peripheral nerves if they are to be useful for repairing damaged nervous tissue. (Preliminary results, which show that they do not elicit a significant adverse cellular reaction when introduced into brain, provide cause for optimism; T. Holmes and colleagues, personal communi-

cation.) Beyond that, we now know of numerous receptors and ligands that mediate neuronal attachment, survival, nerve-fibre outgrowth, synapse formation and modification of synaptic strength. The incorporation of such molecular cues into the peptide scaffolds would be necessary to add further specificity for the attachment and directed outgrowth of particular neurons and nerve fibres. Finally, thorough animal trials with different types of acute and chronic neural damage would have to be carried out before even contemplating trying this approach in humans.

Clearly, the aim of repairing a damaged nervous system remains a long way off. But the further development of biological materials, and the possibility of combining them with cell-based therapies, may bring us closer to realizing that goal. ■

Melitta Schachner is in the Zentrum für Molekulare Neurobiologie, University of Hamburg, Falkenried 94, 20251 Hamburg, Germany.

e-mail: melitta.schachner@zmnh.uni-hamburg.de

1. Holmes, T. C. et al. *Proc. Natl Acad. Sci. USA* **97**, 6728–6733 (2000).

2. Zhang, S., Holmes, T. C., Lockshin, C. & Rich, A. *Proc. Natl Acad. Sci. USA* **90**, 3334–3338 (1993).

3. Zhang, S. et al. *Biomaterials* **16**, 1385–1393 (1995).

Mathematics

Curves and numbers

Ivar Ekeland

The most resounding achievement of mathematics in the twentieth century may well have been the proof of Fermat's

last theorem. As readers of *Nature* may be aware¹, Pierre de Fermat stated this result in the margin of a treatise by the Greek

mathematician Diophantus, along with the remark that he would write the proof somewhere else, because the margin was too small to contain it. That proof has been missing for three and a half centuries, if it ever existed, and in 1993 Andrew Wiles finally supplied a different one, using methods way beyond anything Fermat could have known². Wiles' solution involved a partial proof of another difficult problem, known as the Shimura–Taniyama–Weil conjecture, and this conjecture has finally been proved in full by Christophe Breuil, Brian Conrad, Fred Diamond and Richard Taylor³.

Fermat's theorem states that, for any $n \geq 3$, the only integer solutions of the equation $a^n + b^n = c^n$ are the obvious ones, where a , b or c are 0. For $n = 1$ there are, of course, many other solutions, and for $n = 2$ we have the solution $a = 3$, $b = 4$ and $c = 5$ (a rectangle triangle with integer sides). That such solutions don't exist for higher values of n is surprising, but it is tempting to think that the problem would yield to a bit of smart algebraic manipulation. It does not: generations of amateur mathematicians have tried to do it this way, and failed.

The breakthrough came from another direction: the theory of elliptic curves. These are curves in two variables, x and y , defined by a cubic equation of the form $y^2 = x^3 + px + q$. They can be plotted as points in the x – y plane (Fig. 1), and they have remarkable properties that have captivated mathematicians for centuries. Let me quote just one of them: an elliptic curve has a natural group structure. Group theory is the mathematics of symmetry, and a mathematical group is just a set of elements that can be combined together in pairs to yield another element of the set.

For an elliptic curve (Fig. 1), this group operation is defined geometrically. Given two points A and B on the curve connected by a straight line, the straight line will intersect the curve at a third point C (because the equation defining the curve has degree three). If the curve is reflected about the x -axis, the curve itself remains unchanged, but the point C is changed to point D. As a function of A and B, point D obeys a group law that requires operations to be associative — that is, $A(BD) = (AB)D$. There are two other algebraic laws that define group structure: the existence of the inverse of any element, and the existence of an identity element. Both of these are true of elliptic curves.

Fifty years ago, Shimura, Taniyama and Weil conjectured that all elliptic curves would be 'modular' — a property described in Box 1. In mathematics, a conjecture is something that is believed to be true but is still waiting for a proof. Interest in the Shimura–Taniyama–Weil conjecture was rekindled in 1996, when Gerhard Frey and Ken Ribet observed that if integers a , b and c

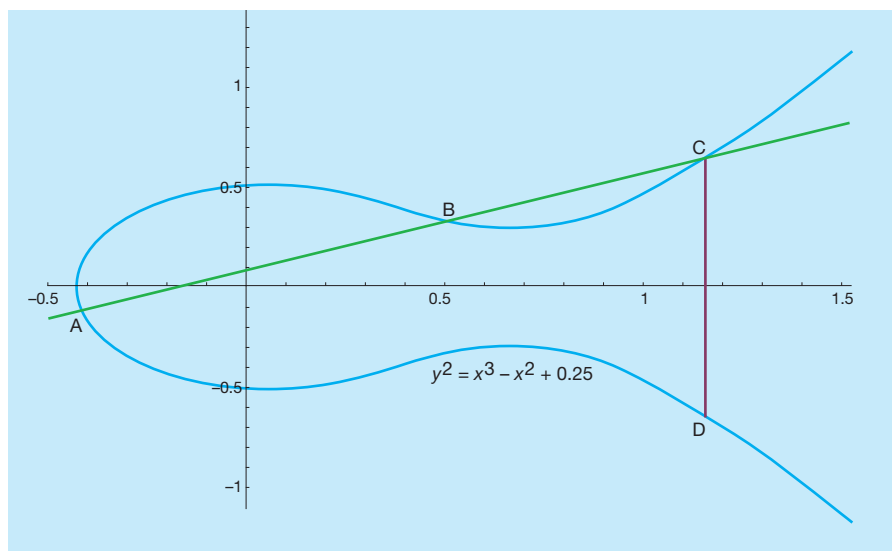


Figure 1 An elliptic curve of the form $y^2 = x^3 - x^2 + 0.25$. What makes elliptic curves worth studying is their natural group structure. The group operation follows from connecting points A and B by a straight line, projecting the line to find a third point C, and then reflecting this point about the x -axis to give point D. Andrew Wiles proved Fermat's last theorem by studying a particular elliptic curve, which involved a partial proof of the Shimura–Taniyama–Weil conjecture. This conjecture has been proved in full by Breuil, Conrad, Diamond and Taylor^{3,4}.

Box 1 Elliptic curves and integer sequences

What does it mean for an elliptic curve to be modular? Given an elliptic curve, $y^2 = x^3 + px + q$, we wish to know if it has integer solutions, that is, points with integer coordinates, and if so, how many. This is a very complicated problem, but it can be simplified by replacing the plane with a periodic lattice.

This is done as follows. Choose a prime number, k , and associate with each integer x the remainder x_k of its division by k (so that $x_k = 0$ if x is a multiple of k). Each point in the plane with integer coordinates

(x, y) now has coordinates (x_k, y_k) : in this way, we reduce the whole plane to a $k \times k$ periodic lattice, containing k^2 points only.

We can look at the same equation, $y_k^2 = x_k^3 + px_k + q$, in the lattice. Let us denote by $n(k)$ the number of solutions of this equation, as a function of k . For instance, if the elliptic curve under consideration is $y^2 = x^3 - x^2 + 1/4$, we get the following numbers: $n(2) = 4$, $n(3) = 4$, $n(5) = 4$, $n(7) = 9, \dots$, $n(10,007) = 9,989, \dots$

Clearly the solutions of the equation in the lattice are related to integer solutions of

the equation in the plane, so that the infinite sequence $n(k)$ encodes some arithmetic information on the elliptic curve under consideration.

The same information is encoded into the Fourier coefficients of an analytic function $f(z)$, which can be stated explicitly as an infinite product, but will not be given here. This function is modular, meaning that it has some special arithmetic properties. Saying that all elliptic curves are modular means that every one of them is related in this way to a modular function. **LE**

solved Fermat's equation $a^n + b^n = c^n$, then the curve $y^2 = x(x - a^n)(x - b^n)$ would be elliptic, and yet would not be modular. What Wiles did was to prove the Shimura–Taniyama–Weil conjecture for the special case that applied to the Frey–Ribet curve. This was enough to prove Fermat's theorem, and brought Wiles immediate fame, but the general case was still unproved. This is what Breuil, Conrad, Diamond and Taylor have achieved^{3,4}.

Theirs is a remarkable proof, because it connects two seemingly unrelated objects: elliptic curves on the one hand, and functions of a complex variable on the other. Each curve can be associated with a sequence of integers (Box 1), whereas each function is associated with another sequence of integers

(basically its Fourier coefficients). Breuil, Conrad, Diamond and Taylor have shown that for every elliptic curve a special function can be found that has the same sequence. Their success has mathematical rewards. For instance, it is now easy to prove generalizations of Fermat's last theorem. But there may be other unexpected benefits: elliptic curves are widely used for public encryption keys, and we now have a new approach to these mathematical objects. Bankers beware! ■

Ivar Ekeland is at the Institut de Finance, Université Paris-Dauphine, 75775 Paris Cedex 16, France.

e-mail: ivar.ekeland@dauphine.fr

1. Swinnerton-Dwyer, P. *Nature* **364**, 13–14 (1993).
2. Wiles, A. *Ann. Math.* **141**, 443–551 (1995).
3. Darmon, H. *Not. Am. Math. Soc.* **46**, 1397–1401 (1999).
4. Conrad, B., Diamond, F. & Taylor, R. *J. Am. Math. Soc.* **12**, 521–567 (1999).

Cell biology

GTPase traffic control

Channing J. Der and William E. Balch

Each small GTP-hydrolysing enzyme (GTPase) of the Rho family regulates a huge variety of cellular processes. This implies that each protein can interact with many different downstream targets that bring about such processes. Indeed, the discovery of such effectors for the Rho proteins continues apace. So the description by Wu and colleagues¹ (page 800 of this issue) of yet another candidate effector for Cdc42 (Fig. 1, overleaf), one of the Rho-family proteins, is not too surprising. What is unexpected is that this effector is a well-characterized subunit of a complex involved in the formation of membrane-clad intracellular vehicles called vesicles. Even more surprising, this subunit may be one of the hitherto-elusive

effectors required for Cdc42 to induce cells to switch to uncontrolled, malignant growth.

Rho-family proteins make up just one branch of the Ras 'superfamily' of small GTPases². All such proteins function as binary switches, which are turned on and off by being bound to GTP or GDP, respectively. But each branch was originally thought to regulate distinct cellular functions. Members of the Rho branch, for example, control organization of the actin-based cytoskeleton, whereas members of the Ras branch regulate cell proliferation. Other branches, such as that occupied by the Arf protein, regulate specific aspects of vesicle movement between organelles.

But the distinctions between the func-



100 YEARS AGO

A novel way of making building land is being carried out not far from New York. The rapidly growing population of this city has made ground scarce on which to build villas and houses for the summer resort of the inhabitants; but the enterprise of the American builder is equal to the emergency, and land is now being literally pumped up from the sea, on which it is intended to erect houses, and to create a new suburb. The site chosen for this venture is the Nassau Beach, on the shore of Jamaica Bay, in Long Island, not far from Brooklyn. The salt marshes bordering on this coast, which for centuries have been overflowed by the tides, and which, of course, while in this condition were utterly unfit for building purposes, are being raised from four to six feet above high water by pumping up the sand, shell and gravel which form the floor of the bay, and delivering this on to the land to be reclaimed... Ten acres have thus been raised since the pumping began a few months ago. A raised road and promenade two miles long and seventy feet wide, and an electric railway, will connect this new suburb with the railway to Brooklyn and New York.

From *Nature* 14 June 1900.

50 YEARS AGO

Simulium damnosum Theobald is a common biting fly in many parts of Africa and an important vector of human onchocerciasis. After the female has bitten a man, the blood in the hinder part of the insect's mid-intestine forms an approximately spherical mass, which retains its form when removed from the gut. During recent dissections, it was noticed that the blood was completely enclosed in a membrane which could not be detected in unfed flies. It was evidently produced by the wall of the mid-intestine during the blood-meal, and appeared to be a peritrophic membrane rather unlike those found in various other insects... When *S. damnosum* is dissected after biting a man and taking up many microfilariae, most of the worms can usually be seen imprisoned by the peritrophic membrane in the mid-intestine and eventually die there. Comparatively few appear in the thoracic muscles of one of these flies and continue development. Frequently, therefore, the membrane protects the fly itself from heavy infection without preventing it from transmitting the parasite.

From *Nature* 17 June 1950.

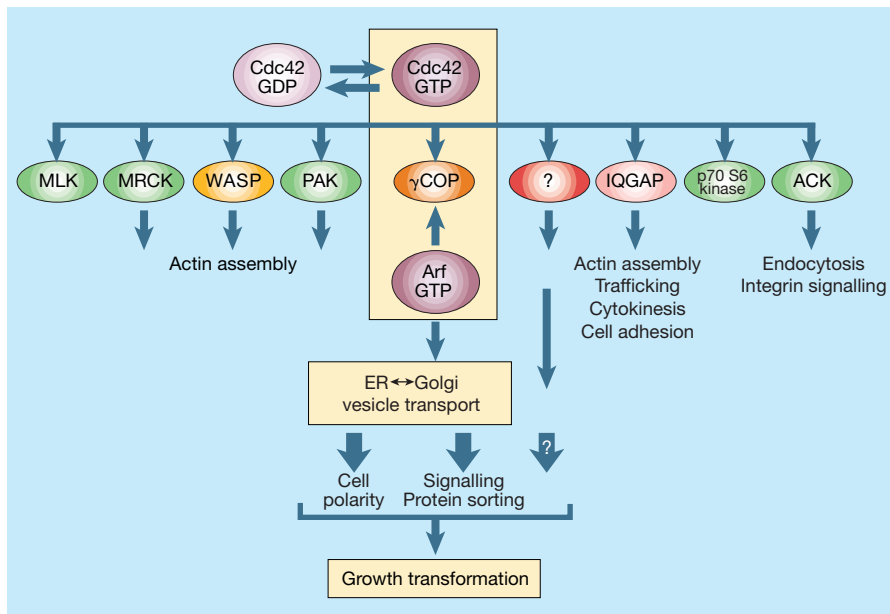


Figure 1 Diverse targets for Cdc42. The activated, GTP-bound form of Cdc42 interacts with a variety of effectors. These include serine/threonine protein kinases (shown in green), IQGAPs and WASP-family proteins^{2,4}. The processes affected by Cdc42-mediated regulation of these proteins are indicated. Wu *et al.*¹ now show that γ COP is also an effector of Cdc42. γ COP is a component of the COPI coatamer protein complex⁵. Like GTP-bound Arf, another GTPase, GTP-bound Cdc42 associates with COPI and regulates vesicle trafficking. Cdc42 and Arf may cooperate to regulate coatamer-mediated vesicle trafficking, or they may regulate distinct mechanisms of trafficking. Surprisingly, γ COP is also required for Cdc42 to induce cellular growth transformation. The cycling of Cdc42 between a GTP-bound (active) and a GDP-bound (inactive) form may be important in regulating the function of γ COP and other effectors (indicated by a question mark) involved in growth transformation.

tions of the various branches have recently begun to blur. Rho proteins in particular take on many tasks. The aberrant activation of these proteins can induce 'growth transformation' — in other words, it can induce cells to become cancerous³ — as can members of the Ras branch. Rho-family proteins may also be involved in regulating the movement of vesicles to and from the plasma membrane^{2,4}. Now, the boundaries become still less clear, as a Rho-family protein, Cdc42, and Arf can both regulate vesicle traffic

through the Golgi by controlling the same effector¹ (Fig. 1). Remarkably, this effector may also account, at least in part, for the growth-transforming activity of Cdc42.

Wu *et al.*¹ use a constitutively activated, GTP-bound, mutant Cdc42 (called Q61L, to denote that glutamine residue 61 is mutated to leucine) to fish for effectors of Cdc42, making the standard assumption that such effectors bind to Cdc42 in its active form. Their search led them to the γ -subunit (γ COP) of COPI, a cytoplasmic 'coatamer'

protein complex⁵ — so called because it coats membrane regions that will bud and form vesicles. Assembly of this complex onto membranes is controlled by Arf⁶. Interactions between Arf and COPI direct the formation of vesicles that are involved in the selective transport of proteins from the Golgi complex back to the endoplasmic reticulum (ER)⁵ and in protein and lipid recycling from the plasma membrane through endosomes⁶. The identification of γ COP as an effector for Cdc42 is consistent with the

Genetics

Fungal get-together

Mushrooms are the tips of icebergs. They are the fruit bodies of basidiomycete fungi that proliferate in soil, extracting nutrients from wood and other solid materials through networks of invasive filaments. Mushroom development usually follows the fusion of two genetically compatible colonies (or mycelia). But research carried out by Robert B. Peabody and colleagues (*Fungal Genet. Biol.* **29**, 72–80; 2000) provides a different picture. It seems that some mushrooms are mosaics that develop from several genetically distinct populations of cells. The mosaics form either through meiosis in the mated mycelium and assortment of haploid strains within the fruit body, or by the copulation of several partners.

Peabody *et al.* looked at a species called *Armillaria gallica* in Massachusetts (a cluster of fruit bodies is shown in the picture). This



is the fungus that became famous when a 10,000-kg mycelium was discovered that has remained genetically stable for more than 1,500 years (M. L. Smith *et al.* *Nature* **356**, 428–431; 1992). Peabody and co-workers analysed tissue samples from mushrooms collected over the past 20 years, and showed that at least two

mycelia had combined to form all mushrooms collected before 1988. They then went on to examine single cells isolated from fruit bodies, the results revealing that some mushrooms contained up to nine genetically distinct individuals.

Although there is strong evidence for mosaicism in specimens collected before 1988, the phenomenon seems to have ended then. The reason is unknown, but it could be due to environmental influences. For example, dry years are less favourable for fungal growth, and in these circumstances the involvement of numerous mycelia might be necessary to support the emergence of fruit bodies.

The identification of mosaicism in *A. gallica* adds a layer of complexity to our understanding of basidiomycete development. Individuals that compete for food in

their mycelial phase must intertwine and become sheathed in a common gel-like extracellular matrix to sculpt the tissues of the mosaic mushroom's stalk, cap and the gills that line the cap. Spore formation will then take place after nuclear fusion and meiosis within specialized cells in the gills called basidia.

Further work will be needed to establish the degree of cooperation between individuals that is required to form mosaics. For instance, determining the proportion of spores that reflect the genotype of each participant should reveal any competition for occupancy of the gill surfaces, and also show who fertilized whom. **Nicholas P. Money**

Nicholas P. Money is in the Department of Botany, Miami University, Oxford, Ohio 45056, USA.
e-mail: moneynp@muohio.edu

authors' previous discovery that Cdc42 can be detected on Golgi compartments⁷. So the Golgi may host a common target for the activities of two distinct small GTPases — Arf and Cdc42.

What, then, might be the function of Cdc42 in vesicle trafficking? Wu *et al.*¹ tackle this problem by following the effects of mutant Cdc42 on the transport of vesicular stomatitis virus G glycoprotein, which is commonly used in vesicle-trafficking studies. Modest expression of so-called 'dominant-negative' Cdc42 mutants⁶ has shown that such mutants interfere with the movement of this viral protein from late Golgi compartments to the basal membrane in epithelial cells. The mutants also affect the ability of the endosomal pathway to maintain the top-to-bottom polarity of epithelial cells.

Wu *et al.*, by contrast, express mutant Cdc42 at higher levels that may stoichiometrically bind the endogenous COPI pool. They find that constitutively GTP-bound Cdc42 (mutant Q61L) and a constitutively GDP-bound form (in which serine 17 is mutated to asparagine) block transport of the viral protein from the ER to Golgi. Given that neither Arf nor COPI is found on ER membranes or is involved in protein export from the ER, it seems that a block in the recycling of proteins from the Golgi to the ER has an indirect, inhibitory effect on movement in the opposite direction, as one cannot proceed without the other. Other results support the idea that the different Cdc42 mutants link to COPI function by triggering a partial collapse of the Golgi into the ER, a COPI-sensitive step. Usually, the recycling of proteins from the Golgi to the ER means that there is a lag in the processing of glycoproteins by enzymes found only in late Golgi compartments. When the Cdc42 mutants are used, this lag does not occur, suggesting that the ER and Golgi compartments are mixed together.

One particular Cdc42 mutant, in which phenylalanine 28 is mutated to leucine, can cause growth transformation of NIH 3T3 cells⁸. As this mutant interferes with the recruitment of COPI to membranes⁵, Wu *et al.* reason that the association of Cdc42 with γ COP may be linked to growth transformation. Indeed, they find that a derivative of this mutant that can no longer bind to γ COP or affect the interaction of the coatomer with the Golgi (or perhaps endosomal compartments) is also unable to cause growth transformation. So interaction with γ COP must be necessary for growth transformation by Cdc42. However, Cdc42 also contains a unique insert sequence that is required for transformation, and Cdc42 lacking this insert still retains the ability to bind γ COP and promote vesicle trafficking. This insert may bind to yet another effector, required together with γ COP to mediate transformation by Cdc42.

A cautionary note must be sounded by the fact that Wu *et al.* used high levels of exogenous, mutated Cdc42 protein. Nonetheless, their results beg the question of how the binding of endogenous Cdc42 to COPI might achieve cellular transformation. Dominant-negative Cdc42 disrupts cell polarity by interfering with vesicle trafficking through the Golgi⁶. So perhaps this loss of cell polarity, in response to dysfunctional control of COPI function by Cdc42, is an important step in the transformation process. One consequence might be the deregulation of the signalling pathways that maintain cells in a dormant state when they are in contact with other cells. The transport of signalling proteins that are normally required for cell maintenance may also become uncontrolled in response to altered trafficking pathways to and from the Golgi, supporting cell proliferation⁹.

It is clear that cells must integrate forwards and reverse vesicle transport, together with the regulation of their cytoskeletons, to control their polarity (or migration) and proliferative behaviour. It seems that when a protein family plays together, cell behaviour stays together. ■

Channing J. Der is in the Department of Pharmacology, University of North Carolina at Chapel Hill, Lineberger Comprehensive Cancer Center, Chapel Hill, North Carolina 27599, USA. e-mail: cjder@med.unc.edu

William E. Balch is in the Department of Cell Biology, Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.

e-mail: webalch@scripps.edu

1. Wu, W. J., Erickson, J. W., Lin, R. & Cerione, R. A. *Nature* **405**, 800–804 (2000).
2. Van Aelst, L. & D'Souza-Schorey, C. *Genes Dev.* **11**, 2295–2322 (1997).
3. Zohn, I. M., Campbell, S. L., Khosravi-Far, R., Rossman, K. L. & Der, C. J. *Oncogene* **17**, 1415–1438 (1998).
4. Ellis, S. & Mellor, H. *Trends Cell Biol.* **10**, 85–88 (2000).
5. Lowe, M. & Kreis, T. E. *Biochim. Biophys. Acta* **1404**, 53–66 (1998).
6. Goldberg, J. *Cell* **100**, 671–679 (2000).
7. Erickson, J. W., Zhang, C.-J., Kahn, R. A., Evans, T. & Cerione, R. A. *J. Biol. Chem.* **271**, 26850–26854 (1996).
8. Lin, R., Bagrodia, S., Cerione, R. A. & Manor, D. *Curr. Biol.* **7**, 794–797 (1997).
9. Allan, B. B. & Balch, W. E. *Science* **285**, 63–66 (1999).

Errata

In the article "Electrons in the looking glass" (Eric Heller *Nature* **403**, 489–491; 2000), the false impression was given that Madhavan *et al.* (*Science* **280**, 567–569; 1998) were the first group to publish STM evidence of Kondo resonances. In fact, Li *et al.* (*Phys. Rev. Lett.* **80**, 2893–2896; 1998) published their results a month earlier. Both groups acknowledge being aware of the other's results before publication.

In Alan P. Boss's article "Three's a crowd" (*Nature* **405**, 405–407; 2000), the source L1551 IRS5 was said to be a member of a triple system, like T Tauri. The text should have been corrected to state that L1551 IRS5 is known to be a binary protostar, whereas T Tauri itself may be one of a system of four stars.

Daedalus

Eye contact

A modern state needs to identify its citizens constantly. Many of them are a potential threat — benefit fraudsters, terrorists, drug pushers, bogus asylum-seekers, élitists, smokers. The citizens even need to identify themselves, both to the authorities and, in their financial dealings, to each other. Yet identity cards are easily stolen, fingerprinting is messy, DNA sampling is intrusive, and none of them work at a distance. Daedalus now has a new idea.

Portraits taken with small cheap cameras are often spoiled by 'red-eye'. The light from the flash enters the eyes of the subject, is reflected from the retinal blood-vessels, and makes his or her pupils appear red in the photograph. What a splendid way, says Daedalus, of obtaining the blood spectrum of a distant subject! DREADCO opticians are devising a camera to maximize the red-eye effect, and to record its spectrum over all the wavelengths that can enter the eye, from near ultraviolet to near infrared. One team is combining a frequency-swept flash with time-resolved charge-coupled-device imaging; another favours a broad-band flash source and a dispersive or Fourier-transform spectrometric detector. Meanwhile, the company's biochemists are exploring the individuality of a blood spectrum.

For a start, it should encode detailed blood-group data: not merely A, B and O, but all the dozens of lesser blood antigens. The plasma polysaccharides and proteins will also tell their story, partly hereditary and partly medical. Indeed, direct indices of criminality, such as metabolites of alcohol, cocaine or nicotine, should show up usefully in the blood spectrum. While not as specific as DNA analysis, red-eye spectrophotometry should still be a powerful identifier. The blood spectrum of every citizen will rapidly be acquired from normal passport or identification-card photography. The resulting database will transform surveillance.

Motorists in speeding cars, rioters in the street, burglars entering protected premises, all will be literally identified in a flash. In daylight the flash might not even be noticed. Counter-measures are possible; but a modern data-dictatorship is used to outflanking them. Just as the police stop any car without a licence plate, and the British government, in its plans to intercept the whole nation's e-mail, will demand that we decrypt for it any message it can't manage to decrypt for itself, so the authorities will arrest anyone wearing dark green glasses.

David Jones

Survival of the Irish elk into the Holocene

Giant deer on the Isle of Man around 9,000 years ago may have been the last of the line.

The giant deer *Megaloceros giganteus* was a celebrated victim of the Late Pleistocene megafaunal extinction, the timing and causes of which are hotly debated¹. Until now, it was believed that the giant deer's demise occurred during the Late Glacial (about 10,600 years ago), before the Pleistocene–Holocene boundary. Here we report new radiocarbon dates from two specimens in stratified contexts, which indicate that a giant deer population still existed around the northern Irish Sea Basin in the early Holocene — 1,400 years after their supposed extinction.

One of the most characteristic animals of the Late Pleistocene European megafauna, the 'Irish elk' had a shoulder height of 2.1 metres and an antler span of up to 3.6 metres. In Ireland and the Isle of Man, most specimens are from deposits of the Allerød Interstadial, about 12,000–11,000 years before present (BP)^{2,3}. However, its distribution extended across Europe to Central Asia, and its fossil record extends back 400,000 years⁴.

An almost complete skeleton of an adult male giant deer was found in 1819 in marl sediments at Loughan Ruy, a basin on the Ballaugh gravel fan, Isle of Man (Fig. 1a)^{5,6}. Another nearly complete adult male skeleton was recovered in 1887 from a marl pit at Close-y-Garey⁷. The Manx Museum in Douglas on the Isle of Man has this on display with a fragment of antler (no. 1977-17) found in 1976 at Glen Wyllin in one of the kettle-holes exposed on the cliffs⁸.

We obtained radiocarbon accelerator mass spectroscopy (AMS) dates for all three specimens: Close-y-Garey, 11,495 ± 95 yr BP within the Allerød interstadial (Lab. no. AA-30361, $\delta^{13}\text{C}\text{‰} = -21.0$); Glen Wyllin, 10,780 ± 95 yr BP during the Younger Dryas (AA-30362, $\delta^{13}\text{C}\text{‰} = -21.2$); and Ballaugh, 9,225 ± 85 yr BP in the early Holocene (AA-29744, $\delta^{13}\text{C}\text{‰} = -20.3$). The AMS dates (quoted uncalibrated) agree with the stratigraphic context of each specimen. The specimen from Ballaugh is the latest reported ¹⁴C date for giant deer anywhere in the world. Another AMS date of 9,430 ± 65 yr BP (AA-18513, $\delta^{13}\text{C}\text{‰} = -21.4$) from a fragmentary *Megaloceros* antler found in the River Cree in southwest Scotland (specimen no. NMSZ 1993.160.12) confirms the survival of the species into the early Holocene in this area.

Equally interesting are the very unusual size and proportions of the Close-y-Garey, and especially the Ballaugh, skeletons. We compared their dimensions to a large sample of male *M. giganteus* from the Irish Late Glacial (Fig. 1b). In all postcranial and

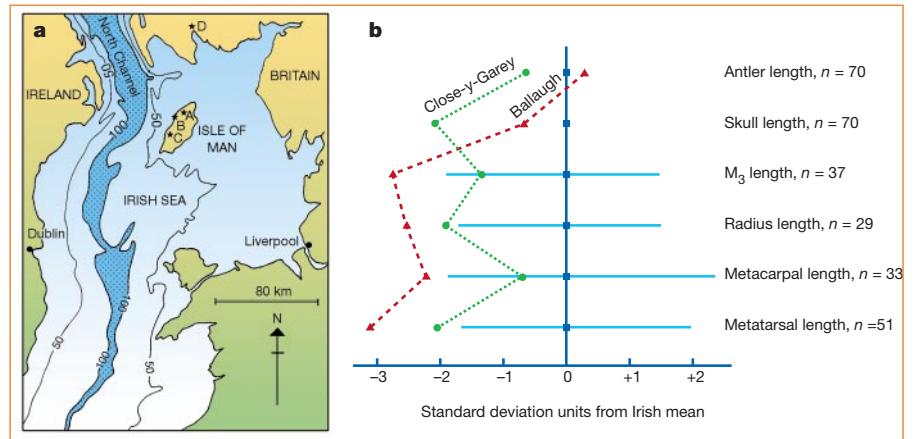


Figure 1 Localities and size comparison of *Megaloceros giganteus* specimens. **a**, Localities (stars) with giant deer remains in the Irish Sea Basin with new radiocarbon dates. A, Ballaugh, Isle of Man; B, Glen Wyllin, Isle of Man; C, Close-y-Garey, Isle of Man; D, River Cree, southwest Scotland. The bathymetry is also shown in metres. The North Channel has been fully open since 12,000 years BP; the shoreline for this period corresponds roughly to the 50-m curve. The Isle of Man became isolated from mainland Britain around 10,000 radiocarbon years BP (ref. 10). **b**, Size of the Ballaugh and Close-y-Garey *M. giganteus* skeletons in comparison with a sample from the Late Glacial of Ireland. Squares, Irish mean; horizontal lines, Irish range; triangles and dashed line, Ballaugh skeleton (National Museums of Scotland, specimen no. NMSZ 1820.16); circles and dotted line, Close-y-Garey skeleton (Manx Museum, no. 1900-1). Data are plotted in units of Irish standard deviations. Measurements of the Irish comparative sample were taken from ref. 4, except antler and skull lengths, which were from ref. 11. Skulls with antlers are known males. Limb bones are isolated specimens but show clear bimodal size distributions⁴, allowing males to be reliably separated; the third lower molar (M₃) shows little or no sexual dimorphism so the total sample was used. Antler length is direct length from the rose to the distal end of the palm; skull length is condylobasal length; M₃ length is taken at crown base lingually; limb bone lengths are the longest dimension.

dental dimensions, the Ballaugh skeleton is below the entire Irish range and more than two standard deviations from its mean (significant at 95% level). This indicates that size decreased into the Holocene. The Allerød Close-y-Garey skeleton is somewhat larger: within the Irish range on two measurements, below it on two others.

Whereas the Close-y-Garey skeleton has relatively similar proportions to the contemporaneous Irish sample, the early Holocene Ballaugh animal has a very large skull compared with its postcrania (Fig. 1b). Exaggerated head-to-body size ratio is common in dwarfed populations of mammal species⁹. It is not clear whether the comparatively small size of the Isle of Man specimens is a geographical or a chronological effect — the Isle of Man became isolated from mainland Britain at about 10,000 yr BP (Fig. 1a)¹⁰.

Despite the small body size, the antlers of both skeletons are well within the Irish size range for adult males (Fig. 1b). There is strong intraspecific allometry among Irish *M. giganteus*¹¹, with smaller animals having relatively shorter antlers, described by $L = 0.00025B^{2.115}$, where B is the basal skull length and L is the length of the antler from the rose to the end of the palm. In the Ballaugh specimen, $B = 468$ mm, predicting an antler length of 1,110 mm; the actual value

is 1,218 mm. In the Close-y-Garey specimen, $B = 450$ mm, predicting a length of 1,022 mm; the actual antler length is 1,090 mm. Reduced display organs are presumed to signify a prelude to extinction from nutritional stress¹², but the Ballaugh and Close-y-Garey individuals show that sexual selection maintained a relatively large antler size until the last documented survival of the species.

Explanations for megafaunal extinctions have generally assumed that these were pre-Holocene and invoke causes from within the Late Pleistocene, whether human or environmental. The nutrient requirements of *M. giganteus* have been modelled, particularly in relation to the substantial costs of growing the huge antlers¹³. Based on the latest date reported, within the Younger Dryas (10,610 ± 495 yr BP (UB-2699), from Ballybetagh, Ireland), extinction is explained in terms of the reduced forage density during that episode, leading to a conflict between sexual selection for increased body and antler size, and nutritional selection pressures for reducing them¹³.

These factors probably reduced populations and possibly the body size of *M. giganteus* in the terminal Pleistocene, but our dating indicates that they cannot have caused its extinction. Survival of *M. giganteus* in the temperate, forested

environment of northwestern Europe in the early Holocene allows the possibility that Mesolithic hunters could have been responsible for the giant deer's final demise, although the terminal dates for *M. giganteus* pre-date the earliest human evidence in both the Isle of Man and Ireland¹⁴. Alternatively, it is possible that the Holocene vegetational zonation was unlike that of previous interglacials and so led to the demise of species adapted to a 'mosaic' environment¹².

It is hard to generalize about causes of extinction as different factors may eliminate populations in different areas. Our new dates for *Megaloceros* add an important new species to the list of Holocene survivors from the Pleistocene world. For the woolly mammoth *Mammuthus primigenius*, the species most extensively dated, recent finds on Wrangel Island in the Siberian Arctic indicated an unexpected survival into the Holocene¹⁵. It is possible that we are witnessing a pattern of terminal survival of Pleistocene megafauna on islands at the margins of continents.

Silvia Gonzalez*, Andrew C. Kitchener†, Adrian M. Lister‡

*School of Biological and Earth Sciences, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK

†Department of Geology and Zoology, National Museums of Scotland, Chambers Street, Edinburgh EH1 1JF, UK

‡Department of Biology, University College London, Gower Street, London WC1E 6BT, UK

e-mail: A.Lister@ucl.ac.uk

1. Stuart, A. J. *Biol. Rev.* **66**, 453–562 (1991).
2. Mitchell, G. F. & Parkes, H. M. *Proc. R. Soc. Lond. B* **258**, 291–314 (1949).
3. Dickson, C. A., Dickson, J. H. P. & Mitchell, G. F. *Phil. Trans. R. Soc. Lond. B* **258**, 31–79 (1970).
4. Lister, A. M. *Zool. J. Linn. Soc.* **112**, 65–100 (1994).
5. Hibbert, S. *Edinb. J. Sci.* **3**, 15–28 (1825).
6. Oswald, H. R. *Edinb. J. Sci.* **3**, 28–31 (1825).
7. Kermod, P. M. C. *Geol. Mag. n.s. decade IV*, **5**, 116–119 (1898).
8. Reeves, G. M. *Proc. IOM Nat. Hist. Antiq. Soc. n.s. VIII* (no. 4, 1978–80), 416–422 (1982).
9. Lister, A. M. *Geol. Soc. Spec. Publ.* **96**, 151–172 (1995).
10. Lambeck, K. J. *Geol. Soc. Lond.* **11**, 437–448 (1995).
11. Gould, S. J. *Evolution* **28**, 191–220 (1974).
12. Guthrie, R. D. in *Quaternary Extinctions: A Prehistoric Revolution* (eds Martin, P. S. & Klein, R. G.) 259–298 (Univ. Arizona Press, Tucson, 1984).
13. Moen, R. A., Pastor, J. & Cohen, Y. *Evol. Ecol. Res.* **1**, 235–249 (1999).
14. Woodman, P. C. *Br. Archaeol. Rep.* no. 58 (1978).
15. Vartanyan, S. L., Garutt, V. E. & Sher, A. V. *Nature* **362**, 337–340 (1993).

Cell signalling

Control of free calcium in plant cell nuclei

Free calcium ions stimulate a huge variety of processes inside the cell¹, eliciting specific responses that depend on their spatio-temporal concentrations^{2,3}. Here we investigate how these changes in calcium concentration are triggered in the cytosol and nucleus of plant cells, and find that they are independently controlled in the two compartments. Our results indicate that in plants some processes in the nucleus may be executed in response to an autonomously regulated nuclear calcium signal, although it is not clear whether this happens in animal cells as well^{4,5}.

To compare nuclear and cytosolic calcium signals in plant cells, we used protoplasts from transgenic tobacco plants that constitutively express the calcium reporter aequorin either in the cytosol⁶ or in the nucleus⁷. The agent mastoparan⁸ stimulated large and transient increases in free calcium in both the cytosol and nucleus in a dose-dependent manner, with the maximum increase in free-calcium concentration being in the micromolar (μM) range for both compartments.

We found, however, that the cytosol and nucleus were differentially sensitive to mastoparan, with half-maximal (in terms of peak height) increases in free calcium in the cytosol and nucleus being elicited by 1 μM

and 5 μM mastoparan, respectively (results not shown). The nuclear compartment consistently responded later than the cytosol, regardless of mastoparan concentration (the delay ranged from 13 ± 2.5 s in the presence of 0.5 μM mastoparan (not shown) to 4 ± 0.5 s with 5 μM mastoparan (Fig. 1a)).

We reasoned that the different response of the two cellular compartments to mastoparan might be due to a difference in the intracellular movement either of calcium itself or of a regulatory compound(s) from the cytosol to the nucleus, or to a specific and independent response by the nucleus itself. To test whether the nucleus could mount an autonomous calcium response to mastoparan, we compared the response to mastoparan of intact protoplasts with that of protoplasts that had been lysed under mild hypotonic-shock conditions⁹ so that their nuclei were exposed.

As expected, 5 μM mastoparan elicited an increase in free calcium in intact protoplasts expressing aequorin in either the cytosol or the nucleus (Fig. 1b, traces 1 and 3). Only a residual response (1–3%, relative to intact protoplasts) to mastoparan was recorded for the lysate prepared from protoplasts expressing aequorin in the cytosol (Fig. 1b, trace 2). In contrast, nuclear aequorin could still respond to mastoparan, with the size of the response being 70% of that by intact protoplasts (Fig. 1b, trace 4).

Under these conditions, addition of 1–10 mM calcium to the lysed nuclear

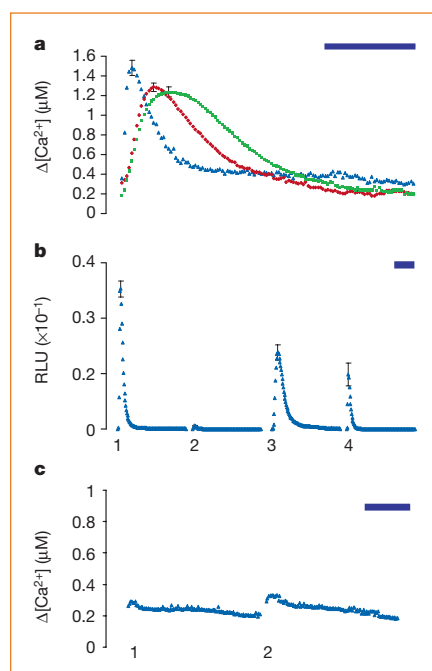


Figure 1 Changes in free calcium concentration in the cytosol and nucleus of tobacco protoplasts. Protoplasts expressing the calcium reporter aequorin in the cytosol or in the nucleus are referred to as 'cyt' or 'nuc', respectively. **a**, Time course of changes in free calcium concentration in intact cyt (blue triangles), intact nuc (green squares) and lysed nuc (red diamonds) protoplasts induced in response to 5 μM mastoparan. **b**, Responses of intact cyt (trace 1), lysed cyt (trace 2) and of intact nuc (trace 3) and lysed nuc (trace 4) protoplasts to 5 μM mastoparan. RLU, relative light unit. **c**, Effect on nuclear Ca^{2+} concentration after addition of 1 mM (trace 1) or 10 mM (trace 2) CaCl_2 to the incubation medium containing lysed nuc protoplasts. Scale bars, 10 s.

preparation had no effect on the existing concentration of free calcium (Fig. 1c), suggesting that the nuclear compartment is not passively permeable to calcium, at least in the form of isolated nuclei. The calcium response to mastoparan recorded from lysed protoplasts expressing aequorin in the nucleus was always earlier than that in the corresponding intact protoplasts in the presence of either 1 μM (not shown) or 5 μM mastoparan (Fig. 1a).

Taken together, our results suggest that nuclei from plant cells are capable of generating their own calcium signals independently of changes in calcium ion concentration in the cytosol.

Nicolas Pauly*, Marc R. Knight†, Patrice Thuleau*, Arnold H. van der Luit‡, Marc Moreau§, Anthony J. Trewavas||, Raoul Ranjeva*, Christian Mazars*

*Signaux et Messages Cellulaires chez les Végétaux, UMR-CNRS/UPS 5546, Pôle de Biotechnologie Végétale, 24 Chemin de Borde Rouge, BP 17 Auzeville, 31326 Castanet-Tolosan, France

†Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

‡Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

§Centre de Biologie du Développement, UMR CNRS/UPS 5547, Université Paul Sabatier, 118

Route de Narbonne, 31062 Toulouse cedex, France
 Institute of Cell and Molecular Biology,
 Mayfield Road, University of Edinburgh,
 Edinburgh EH9 3JH, UK
 e-mail: mazars@cict.fr

1. Berridge, M. J. *Nature* **361**, 315–325 (1993).
2. Clapham, D. E. *Cell* **80**, 259–268 (1995).
3. McAinsh, M. R. & Hetherington, A. M. *Trends Plant Sci.* **3**, 32–36 (1998).
4. Hardingham, G. E., Chawla, S., Johnson, C. M. & Bading, H. *Nature* **385**, 260–265 (1997).
5. Malviya, A. N. & Rogue, P. J. *Cell* **92**, 17–23 (1998).
6. Knight, M. R., Campbell, A. K., Smith, S. M. & Trewavas, A. J. *Nature* **352**, 524–526 (1991).
7. Van der Luit, A. H., Olivari, C., Haley, A., Knight, M. R. & Trewavas, A. J. *Plant Physiol.* **121**, 705–714 (1999).
8. Takahashi, K., Isobe, M. & Muto, S. *Biochim. Biophys. Acta* **1401**, 339–346 (1998).
9. Stoppin, V., Vantard, M., Schmit, A. C. & Lambert, A. M. *Plant Cell* **6**, 1099–1106 (1994).

Cell biology

Risky immortalization by telomerase

Senescence naturally limits the proliferation of mammalian cells in culture, possibly by shortening the telomere regions at the ends of chromosomes during cell division^{1,2}. In support of this idea, introducing TERT, the catalytic subunit of telomerase — the enzyme that maintains chromosome ends — into certain cell types can extend their lifespan and potentially immortalize them^{3,4}. It has been proposed that treatment with exogenous TERT might be useful for cell-based therapies by allowing indefinite expansion of normal human cells without damaging their genomes^{5,6}. But we show here that TERT-driven cell proliferation is not genoprotective because it is associated with activation of the *c-myc* oncogene.

Human mammary epithelial cell (HMEC) cultures normally stop dividing at 55–60 population doublings (PDs). We infected these cells with a human(h)TERT retrovirus at PD40 and maintained them until PD250 (ref. 4), then tested whether telomerase activity was essential for this immortalized phenotype by excising the hTERT retrovirus at PD150 using Cre recombinase⁷. The resulting HMEC–Cre cells were maintained for at least another 20 population doublings and we saw no decline in growth rates in either pooled cells or individual clones. We used Southern blots of genomic DNA to confirm that TERT had been removed (data not shown). To our surprise, telomerase activity remained high compared with the control parental culture, which had undetectable activity (Fig. 1a).

Ectopic expression of *c-myc* activates telomerase in HMECs⁴, and hTERT is a direct transcriptional target of *c-Myc*⁸. To determine whether activation of *c-myc* was

responsible for the telomerase activity found in our HMEC–Cre cultures, we measured the amount of Myc protein and found a two- to threefold increase in *c-Myc* compared with vector control cells (Fig. 1b). This is comparable to the amounts found in a culture immortalized by a *myc*-encoding retrovirus and in a breast-cancer cell line, HBL100 (Fig. 1b). The excision process did not itself cause this increase in *c-Myc*, as expression of *c-myc* was high before excision and also in clone 4, which still retained TERT (Fig. 1b).

We examined *c-myc* expression in HMEC–hTERT at different population doublings to determine when it was up-regulated and found that it increased between 107 and 135 population doublings (Fig. 1c). Also, GADD45 protein, whose expression is repressed by Myc, decreased markedly in HMEC–hTERT at PD135 and in HMEC–Cre at PD150 compared with vector control cells and HMEC–hTERT at PD70 (Fig. 1d). These results indicate that,

under standard culture conditions, extension of lifespan by telomerase selects for *c-myc* overexpression in HMECs.

Activation of the *c-myc* oncogene by overexpression, gene amplification, translocation and possibly mutation occurs in a wide variety of tumour types⁹. We have shown that, although telomerase activation extends the lifespan of HMECs, it is also associated with overexpression of *c-myc* and so is not indefinitely genoprotective (even though the chromosome number in such cells is normal; data not shown). Paradoxically, the extension of lifespan that is conferred by TERT causes *c-myc* activation, and this immortalizes cells, in part by activating TERT expression. Furthermore, in HMEC cultures, TERT expression has little, if any, immortalizing potential until the *p16* tumour-suppressor gene has been inactivated¹⁰. These findings indicate that the use of hTERT for expansion of normal human cells for therapeutic purposes must be approached with caution.

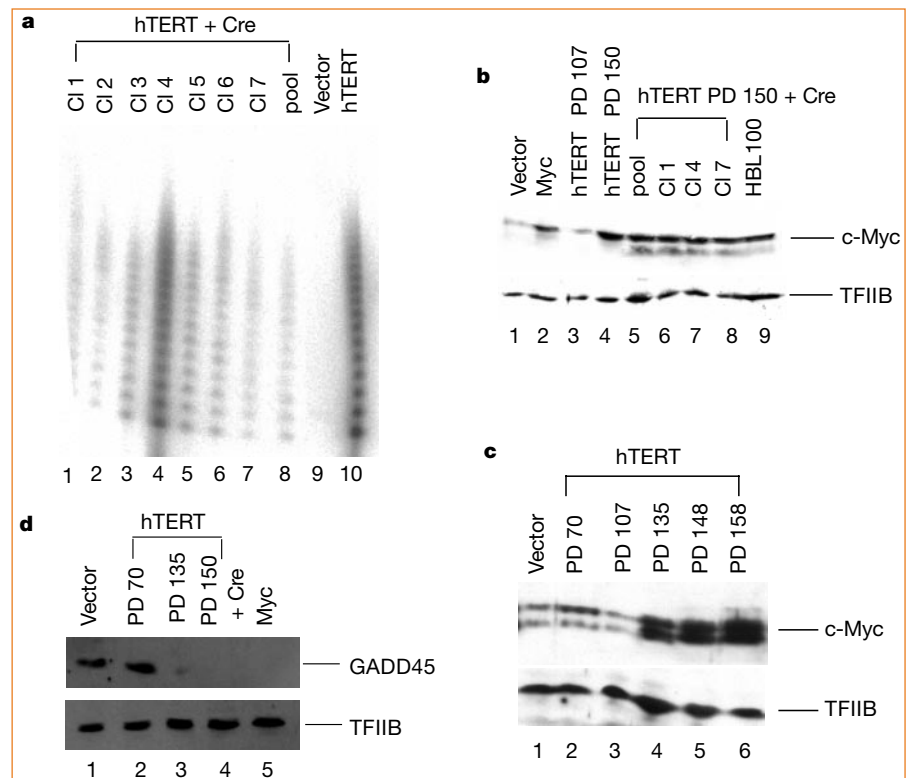


Figure 1 *c-myc* activity is increased in immortalized HMEC–hTERT cells. HMEC 184 spiral K cells were provided by M. Stampfer. All HMEC-derived cultures were maintained in complete mammary epithelium growth medium (MEGM, Clonetics) under standard conditions¹¹. **a**, Telomerase activity in HMEC–Cre cells. We infected HMEC–hTERT cells at PD150 with retroviruses that direct the expression of Cre recombinase, and generated a pool (lane 8) and seven clones (Cl; lanes 1–7). The pool and individual clones, except clone 4, lost the hTERT retroviral cassette. After maintaining cultures for a further 20 population doublings (PDs), we prepared cell lysates for TRAP assays¹². Lane 9, vector control cells at PD40; lane 10, non-excised HMEC–hTERT cells at PD150. Each lane corresponds to 10,000 cells. Results were similar in clones obtained after a second round of subcloning of Cre clone 6. **b**, Immunoblot using rabbit polyclonal anti-cMyc antibody (N-262, Santa Cruz) showing c-Myc in cell lysates from vector-infected HMECs (lane 1), cells immortalized by ectopic expression of c-Myc (lane 2), hTERT-expressing cells at PD107 (lane 3) and PD150 (lane 4), Cre-infected HMEC–hTERT pool (lane 5) and clones 1, 4 and 7 (lanes 6–8), and tumour-cell line HBL100 (lane 9). **c**, Immunoblot showing the amount of c-Myc in cell lysates from vector control cells (lane 1) and HMEC–hTERT cells at different PDs (lanes 2–6). **d**, Immunoblot using rabbit polyclonal GADD45 antibody (H-165, Santa Cruz), showing GADD45 in cell lysates from vector control cells (lane 1), hTERT-expressing cells at PD70 (lane 2) and PD135 (lane 3), Cre-infected HMEC–hTERT at PD150 (lane 4) and cells immortalized by ectopic expression of c-Myc (lane 5). In **b–d**, TFIIB protein was used to normalize loading.

Jing Wang*, Gregory J. Hannon†, David H. Beach‡

*Genetica Inc., 1 Kendall Square, Building 600, Cambridge, Massachusetts 02139, USA

†Cold Spring Harbor Laboratory, PO Box 100, Cold Spring Harbor, New York 11743, USA

‡Wolfson Institute for Biomedical Research, University College London, Gower Street, London WC1B 6BT, UK
e-mail: d.beach@ucl.ac.uk

1. Harley, C. B., Fitcher, A. B. & Greider, C. W. *Nature* **345**, 458–460 (1990).
2. Harley, C. B. *Mutat. Res.* **256**, 271–282 (1991).
3. Bodnar, A. G. et al. *Science* **279**, 349–352 (1998).
4. Wang, J., Xie, L. Y., Allan, S., Beach, D. & Hannon, G. J. *Genes Dev.* **12**, 1769–1774 (1998).
5. Morales, C. P. et al. *Nature Genet.* **21**, 115–118 (1999).
6. Jiang, X. R. et al. *Nature Genet.* **21**, 111–114 (1999).
7. Sun, P., Dong, P., Hannon, G. J. & Beach, D. *Science* **282**, 2270–2272 (1998).
8. Wu, K. J. et al. *Nature Genet.* **21**, 220–224 (1999).
9. Alitalo, K. et al. *Biochim. Biophys. Acta* **907**, 1–32 (1987).
10. Kiyono, T. et al. *Nature* **396**, 84–88 (1998).
11. Foster, S. A. & Galloway, D. A. *Oncogene* **12**, 1773–1779 (1996).
12. Kim, N. W. et al. *Science* **266**, 2011–2015 (1994).

Insect perception

Do cockroaches ‘know’ about fluid dynamics?

Animals use their senses to extract information from the world around them, so they need to be able to gauge the physical properties of their environment in order to build up an accurate perception of it. For example, a bat needs to ‘know’ the velocity of sound to estimate how far away an object is, although input to a sensory system may often exploit more complicated properties than this. Here we measure the response by the wind-sensing system of the American cockroach (*Periplaneta americana*) to a complex hydrodynamic flow. We find that the insect’s interneurons relay crucial information about the wind’s spectral properties, which may warn it of approaching predators.

The cockroach senses minute air movements using tiny hairs on two posterior appendages called cerci¹. It can surmise the direction of an attack and scurry away to avoid being eaten. Neural signals from the hairs converge on the terminal abdominal ganglion where the wind information is processed, and are then conveyed further by giant interneurons. Although this system has many of the properties of more complex systems, it remains simple enough to be tractable for study.

We produced random wind stimuli with defined spectral properties and measured the average firing rates of several interneurons in response to this stimulus. For a given spectral shape, the total power of the stimulus did not change the steady-state firing rates of the interneurons (Fig. 1a).

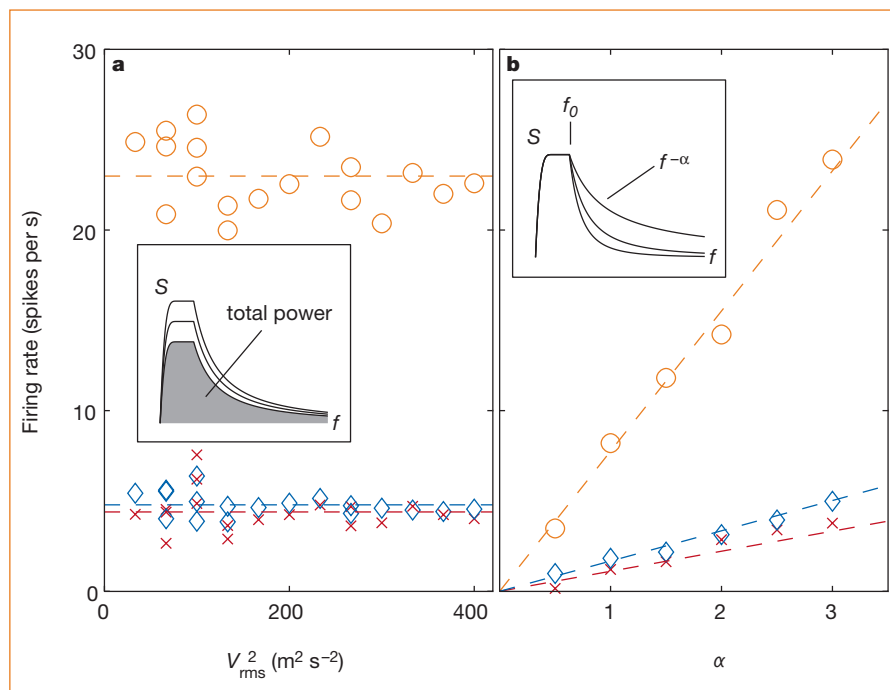


Figure 1 Average firing rate for random wind stimuli with different spectral parameters. Rates are shown for two typical interneurons (red crosses and blue diamonds) and for all interneurons together (orange circles). Insets show spectral density, S , as a function of frequency, f , indicating how the spectral parameters were changed. The frequency f_0 was held constant for these experiments at 10 Hz. The total power, which is directly proportional to the r.m.s. of the square of the wind velocity, V_{rms}^2 , and the high-frequency ‘roll-off’ parameter, α , were changed independently. **a**, Firing rate as a function of total power of wind spectra with $\alpha = 3$. **b**, Firing rate as a function of α shows strong dependence on the extent of the high-frequency tail.

Changing the high-frequency roll-off, on the other hand, strongly influenced the firing rates of all of the cells (Fig. 1b). Thus, exposing the system to narrow-band, low-frequency noise produces a strong cell response — that is, a high firing rate — whereas exposure to wide-band stimuli does not. In the limiting case of white noise, the firing rate is almost zero — in spite of the fact that the afferent neurons are known² to respond to excitations above 100 Hz. Similar effects are expected for this type of stimulus in other systems³.

Let us now consider the typical airflow in a cockroach’s environment. The Reynolds number gives an indication of the degree of turbulence⁴: given the typical size of surrounding objects (less than about 1 m in size) and the relevant wind velocities (0.1 m s^{-1}), the Reynolds number is $\text{Re} \approx 10^3$, so cockroaches live in a world that is often turbulent. Spectra with long, high-frequency tails are characteristic of turbulent airflow⁵. In contrast, the first sign of an approaching predator is slow-moving air, whose spectrum has only low frequencies: in the case of attacking toads and wasps, timescales are typically about 50 ms — corresponding to frequencies below about 20 Hz (refs 6,7). A low-frequency, narrow-bandwidth stimulus may thus be an indicator of a possible attack.

It is evident from Fig. 1 that the average firing rate of the cockroach interneurons conveys information about the spectral

properties of the prevailing air movement, which change when a predator approaches. Thus, the insect’s awareness of these properties and its ability to detect deviations from the norm — in the form of an excess of low-frequency winds — may help it to survive.

D. Rinberg, H. Davidowitz

NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540-6634, USA
e-mail: dima@research.nj.nec.com

1. Camhi, J. M. *Neuroethology* (Sinauer, Sunderland, Massachusetts, 1984).
2. Kondoh, Y., Arima, T., Okuma, J. & Hasegawa, Y. *J. Comp. Physiol. A* **169**, 653–662 (1991).
3. Vedenina, V. Y., Rozkova, G. I., Panjutin, A. K., Byzov, A. L. & Kämpfer, G. *J. Comp. Physiol. A* **183**, 553–561 (1998).
4. Purcell, E. M. *Am. J. Phys.* **45**, 3–11 (1977).
5. Tritton, D. J. *Physical Fluid Dynamics* (Oxford Univ. Press, Oxford, 1998).
6. Plummer, M. R. & Camhi, J. M. *J. Comp. Physiol. A* **142**, 347–357 (1981).
7. Gnatzy, W. & Kämpfer, G. *J. Comp. Physiol. A* **167**, 551–556 (1990).

Erratum

Focusing hard X-rays with old LPs

B. Cederström, R. N. Cahn, M. Danielsson, M. Lundqvist, D. R. Nygren
Nature **404**, 951 (2000)

An editing error altered the intended meaning of the last two sentences of the seventh paragraph, which should read “We used PVC for focusing. As it contains a large fraction of chlorine, it provides less gain than PMMA, for example.” Thus PVC is inferior to PMMA, but we used it for demonstration anyway.

The evolutionarily conserved BMP-binding protein Twisted gastrulation promotes BMP signalling

Michael Oelgeschläger*, Juan Larraín*, Douglas Geissert & Eddy M. De Robertis

Howard Hughes Medical Institute and Department of Biological Chemistry, University of California, Los Angeles, California 90095-1662, USA

* These authors contributed equally to this work

Dorsal–ventral patterning in vertebrate and *Drosophila* embryos requires a conserved system of extracellular proteins to generate a positional information gradient. The components involved include bone morphogenetic proteins (BMP/Dpp), a BMP antagonist (Chordin/Short gastrulation; Chd/Sog) and a secreted metalloproteinase (Xolloid/Tolloid) that cleaves Chd/Sog. Here we describe *Xenopus* Twisted gastrulation (xTsg), another member of this signalling pathway. xTsg is expressed ventrally as part of the BMP-4 synexpression group and encodes a secreted BMP-binding protein that is a BMP signalling agonist. The data suggest a molecular mechanism by which xTsg dislodges latent BMPs bound to Chordin BMP-binding fragments generated by Xolloid cleavage, providing a permissive signal that allows high BMP signalling in the embryo. *Drosophila* Tsg also binds BMPs and is expressed dorsally, supporting the proposal that the dorsal–ventral axis was inverted in the course of animal evolution.

Dorsal–ventral patterning in vertebrates is regulated by a gradient of BMP activity. BMPs are expressed relatively uniformly in a wide area of the gastrulating *Xenopus* embryo and the gradient is thought to be generated by the localized secretion of BMP antagonists, such as Chordin and Noggin, by the dorsal lip or Spemann's organizer^{1,2}. A further level of regulation is introduced by a secreted zinc metalloproteinase, Xolloid, which cleaves inactive Chordin–BMP complexes, resulting in the reactivation of BMP signalling in the embryo^{3,4}. This model of dorsal–ventral patterning has been validated by genetic studies in zebrafish^{5–10}. Chordin contains four cysteine-rich (CR) domains of about 70 amino acids each, and the Xolloid cleavage sites are located at conserved aspartic acid residues just downstream of CR1 and CR3 (refs 3, 11). Individual cysteine-rich domains, in particular CR1 and CR3, bind BMP, albeit with a 10-fold lower affinity than full-length Chordin¹². Microinjection of CR1 or CR3 messenger RNA results in dorsalization and induction of secondary axes in *Xenopus* embryos. Thus, even after cleavage by Xolloid, the Chordin fragments can still inhibit BMP signalling¹². This observation indicated that additional factors might be required to release BMP from the Chordin fragments generated by Xolloid to reactivate BMP signalling through its cognate receptor. In *Drosophila*, seven zygotic genes have been proposed to regulate dorsal–ventral patterning^{1,13}. Among them, *decapentaplegic* (*dpp*) and *screw* (*scw*) encode BMP homologues that promote dorsal cell fates such as amnioserosa and inhibit development of the ventral central nervous system^{13–16}. The *chordin* homologue *short gastrulation* (*sog*) is expressed ventrally and promotes central nervous system development^{17–19}.

The phenotype of *sog* loss-of-function mutants is intriguing: as expected for a Dpp/Scw antagonist, ventral structures are lost but, in addition, the amnioserosa is reduced. This result is paradoxical, as the amnioserosa is the dorsal-most tissue and therefore Sog, a BMP antagonist, is required for maximal BMP signalling^{20–23}. A model proposed to explain the role of Sog in promoting peak Dpp activity suggests that Sog–BMP complexes may permit the diffusion of BMPs originating from more ventral regions, which are then released dorsally by the proteolytic activity of Tolloid²¹. The recent demonstration that BMPs remain bound to individual cysteine-rich domains, which remain intact in the Chordin proteolytic products¹², makes this interpretation unlikely, unless an additional

factor that releases BMP from the cysteine-rich modules is proposed. The *twisted gastrulation* gene encodes a secreted protein that is specifically required for the differentiation of amnioserosa cells in *Drosophila*^{24,25} and is a candidate for such a factor.

Here we show that the *Xenopus* homologue of *twisted gastrulation* (*xTsg*) shares sequence similarities with the cysteine-rich domains of Chordin, and is part of the BMP synexpression group²⁶. Biochemical studies show that xTsg and *Drosophila* dTsg directly bind BMPs with dissociation constants in the low nanomolar range. In microinjection experiments, xTsg mRNA behaves as an agonist of BMP signalling, ventralizing the *Xenopus* embryo. xTsg competes efficiently with CR1 for binding to BMP and can bind full-length Chordin, forming a ternary complex containing Chordin, BMP and xTsg *in vitro*. The dorsalizing activity of CR1 is readily competed by wild-type xTsg, and is greatly potentiated by reducing endogenous xTsg activity. The results indicate that xTsg is involved in dorsal–ventral patterning, permitting peak BMP signalling by antagonizing the residual anti-BMP activity of the cleavage products of Chordin.

xTsg is expressed in ventral-most tissues

We isolated a full-length xTsg complementary DNA by using a human expressed sequence tag (EST) to probe a *Xenopus* gastrula library. The cDNA encodes a protein sharing 41% amino-acid identity with *Drosophila* Tsg (dTsg), 89% identity with the partial human Tsg sequence and 94% identity with a mouse EST. The xTsg sequence contains a signal peptide, as expected for a secreted protein, and two conserved domains containing multiple cysteines at its amino and carboxy termini (Fig. 1a). Whole-mount *in situ* hybridization and polymerase chain reaction with reverse transcription (RT–PCR) showed that abundant xTsg maternal transcripts are distributed throughout the animal half of the embryo during cleavage stages (Fig. 1b, and data not shown). At the late gastrula stage, maternal transcripts decrease and zygotic transcripts appear specifically in the ventral region of the embryo (Fig. 1c). After neurulation, xTsg transcripts surround ventrally the closed blastopore slit and the neural tube (Fig. 1d). At the tailbud stage, xTsg transcripts are detected in the postanal region, heart and dorsal eye (Fig. 1e, f) and closely mimic the expression patterns of BMP-4 and BAMBI^{27,28} (Fig. 1e–j). The postanal region of xTsg expression derives from the ventral-most tissue²⁹ of the gastrula embryo, as

illustrated by the transplantation experiment shown in Fig. 1k and l. We conclude that *xTsg* is part of the *BMP-4* synexpression group²⁶ and is expressed in the ventral pole of the embryo.

xTsg has ventralizing activity

Microinjection of *xTsg* mRNA into each blastomere of the four-cell embryo resulted in the reduction of dorso-anterior structures (Fig. 2a). This phenotype was reminiscent of the *chordin* zebrafish mutant⁵, and of *Xenopus* embryos microinjected with *Xolloid* mRNA^{3,30} or with low doses of *BMP-4* mRNA³¹. Molecular marker analyses showed that dorsal ectoderm (*Sox-2*) and dorsal mesoderm (*MyoD* and *Shh*) are reduced and that ventral tissues (*BMP-4*) are expanded in *xTsg*-injected embryos (Fig. 2b–e). We conclude from these results that *xTsg* has ventralizing activity.

To determine whether *xTsg* functions in the BMP pathway, epistatic analyses were performed by injecting *xTsg* mRNA into a ventral blastomere together with a dominant-negative BMP receptor

(*tBR*) or extracellular BMP antagonists (*noggin*, *chordin* and *CR1*). *xTsg* had no effect on the formation of secondary axes by *tBR*, indicating that it may ventralize the embryo upstream of the BMP receptor (Fig. 3b). To test whether *xTsg* could antagonize the dorsalizing activity of the proteolytic fragments of Chordin, we generated a construct containing the N-terminal CR1 domain and terminating at the Xolloid cleavage site¹¹. *CR1* mRNA induced secondary axes (although at 32-fold higher molar concentrations than those required for full-length *chordin* mRNA, Fig. 3g), which were blocked by co-injection of *xTsg* (Fig. 3h). *Xolloid* mRNA was able to block the activity of full-length *chordin*, as shown previously³, but had no effect on secondary axes induced by the *CR1* construct (Fig. 3f, i). This indicates that *xTsg* efficiently antagonizes high doses of CR1 downstream of Xolloid cleavage. The ability of *xTsg* to inhibit full-length *Chordin* mRNA (Fig. 3d, e) presumably results from the activity of uniformly expressed Xolloid proteases in the early embryo³⁰. That the ventralizing activity of *xTsg* requires Xolloid cleavage is confirmed by the observation that a CR1 construct containing 80 additional amino acids downstream of the first Xolloid site could not be antagonized by co-injection of either *xTsg* or *Xolloid* mRNAs (unpublished observations). The ventralizing activity of *xTsg* appears to be specific for the Chordin/BMP pathway, as inhibition of BMP by *noggin* was not affected by co-injection of *xTsg* mRNA (Fig. 3c). These epistatic studies are consistent with a model in which *xTsg* would ventralize the embryo by antagonizing the residual anti-BMP activity of Chordin proteolytic cleavage products (Fig. 3j).

xTsg is a BMP-binding protein

When the N-terminal domain of *xTsg* was compared with the cysteine-rich domains of Chordin, we noticed sequence similarities (Fig. 4a). As cysteine-rich domains are BMP-binding modules¹², we tested whether epitope-tagged *xTsg* secreted by transfected 293T cells could bind BMP-4. *xTsg* did bind to BMP-4 in solution, and

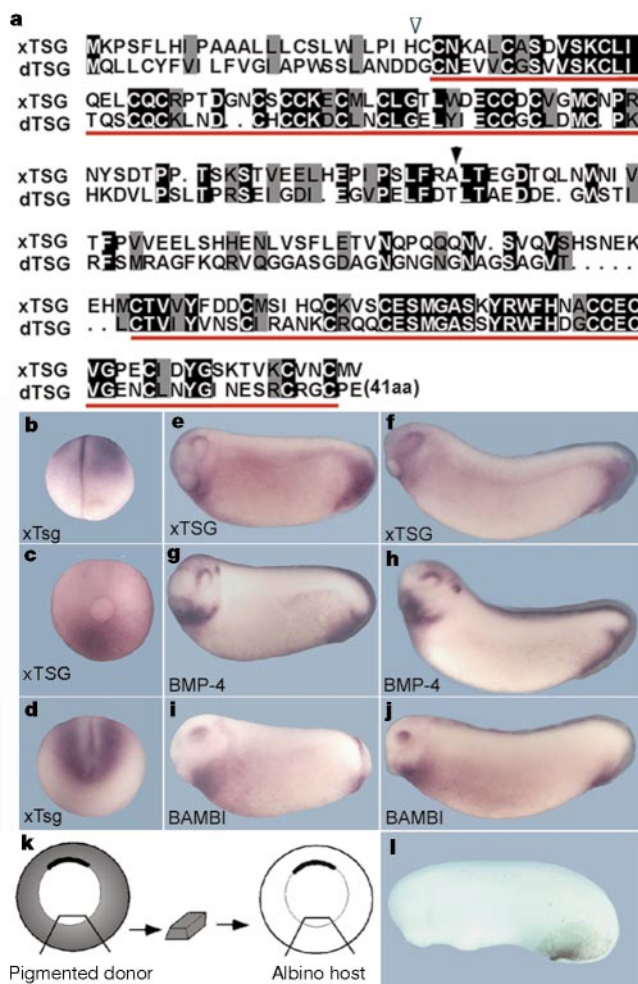


Figure 1 *xTsg* shares two conserved regions with *dTsg* and is co-expressed with *BMP-4* and *BAMBI*. **a**, Alignment of *xTsg* and *dTsg*. The potential cleavage site for the signal peptide (open arrowhead), the conserved regions (red bars) and the position at which the N-terminal and C-terminal fragments were divided (black arrowhead) are indicated. **b–f**, Whole-mount *in situ* analysis of *xTsg* expression. **b**, Four-cell stage; **c**, late gastrula (posterior view); **d**, stage 14. **e**, **f**, At stages 20 and 24, *xTsg* expression marks the postanal region, the heart and the dorsal eye. **g**, **h**, *BMP-4* and **i**, **j**, *BAMBI* expression at stages 20 and 24, respectively. **k**, Experimental design of an isotopic and isochronic transplant of the ventral-most tissue at early gastrula (stage 10½) from a pigmented donor to an albino host ($n = 11$). **l**, The ventral pigmented transplanted tissue populates the perianal region of the early tailbud embryo.

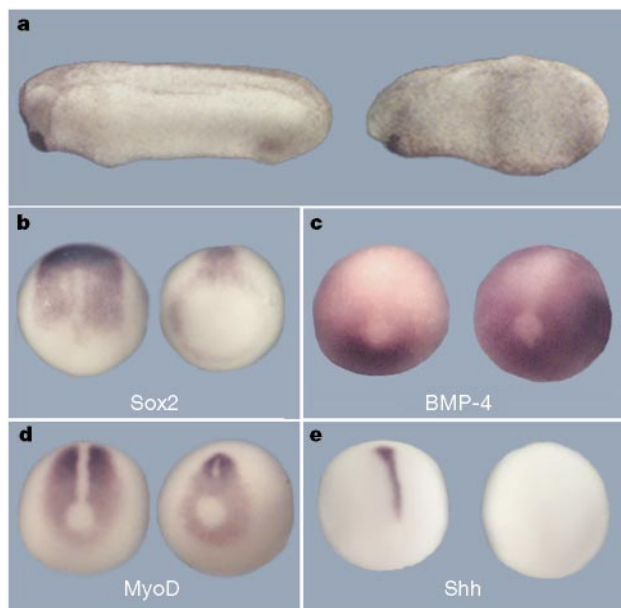


Figure 2 Microinjection of *xTsg* mRNA leads to ventralization of the embryo. **a**, Control embryo (left) and embryo with *xTsg* mRNA microinjected into each animal cell at the eight-cell stage (500 pg total; right), leading to the reduction of dorso-anterior structures. **b–e**, *In situ* analysis of late gastrulae microinjected with *xTsg* mRNA (500 pg per blastomere). Injected embryos (right) and uninjected controls (left) in dorsal view. **b**, *Sox2*; **c**, *BMP-4*; **d**, *MyoD*; **e**, *Shh*. The changes in gene expression are characteristic of ventralization in *Xenopus*³.

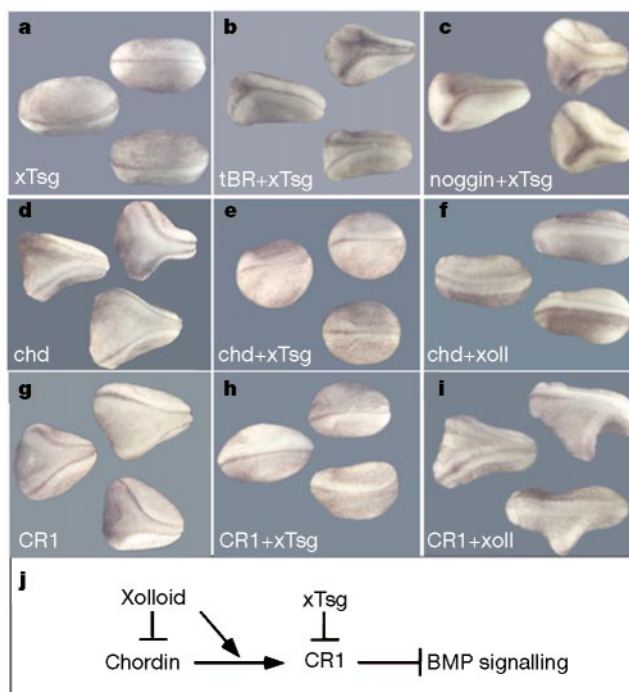


Figure 3 xTsg blocks secondary axis formation by CR1 downstream of Chordin cleavage by Xolloid. **a**, Microinjection of 250 pg *xTsg* mRNA. **b,c**, The induction of secondary axes by 200 pg *tBR* (dominant-negative BMP receptor; **b**) or 5 pg *noggin* mRNA (**c**) is not affected by co-injection of 250 pg *xTsg* mRNA. **d**, Injection of 10 pg *chordin* mRNA induces the formation of secondary axes that are inhibited by co-injection of 250 pg *xTsg* mRNA (**e**) or 200 pg *Xolloid* mRNA (**f**). **g**, Secondary axes induced by injections of 80 pg *CR1* mRNA (a 32-fold molar excess compared with full-length *chordin* mRNA) are

inhibited by co-injection of 250 pg *xTsg* mRNA (**h**) but not by co-injection of 200 pg *Xolloid* mRNA (**i**). Similar results were obtained in at least three independent experiments, with $n = 22-52$ embryos per mRNA combination. All embryos were injected once ventrally at the eight-cell stage. **j**, Summary of the epistatic studies. Xolloid inhibits Chordin activity but not that of CR1, and xTsg can block CR1 function. These epistatic analyses place the xTsg ventralizing activity downstream of Xolloid cleavage and upstream of the receptor.

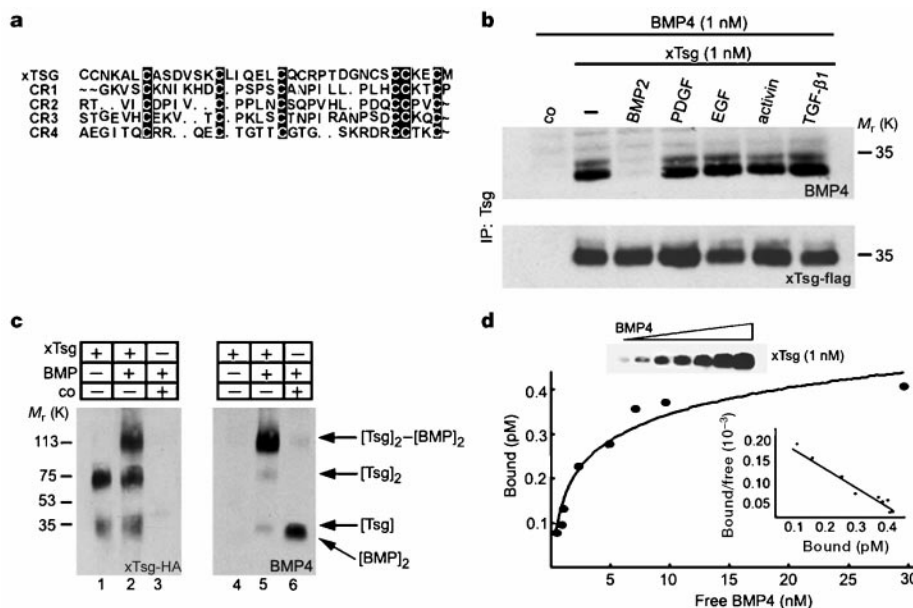


Figure 4 xTsg binds to BMP specifically and directly. **a**, Sequence similarities between Chordin cysteine-rich modules and the N-terminal domain of xTsg. **b**, Western blot analyses of BMP-4 (1.0 nM) bound to xTsg (1.0 nM) after immunoprecipitation by xTsg in the presence or absence of a 10-fold excess of various growth factors. The control (co) lanes in this and Figs 5–7 contain conditioned medium from mock-transfected 293T cells. To measure the recovery of xTsg after immunoprecipitation, the same membrane was stripped and probed for xTsg protein (bottom). **c**, Crosslinking analysis of xTsg–BMP-4 complexes with DSS. Left, immunoblot probed for xTsg. Right, the same membrane

probed for BMP-4 after stripping. **d**, Equilibrium binding of increasing concentrations of BMP-4 (0.2–30 nM) to 1 nM Tsg. Binding was for 3 h at 4 °C, and bound and free BMP-4 were separated by immunoprecipitation; crosslinking agents were not used. The amount of bound BMP-4 was determined as described¹². Each data point was in duplicate, and two independent experiments were performed for each protein. Scatchard analyses of the dTsg data gave an apparent K_d of 2.5 nM. The same results and K_d were obtained for the xTsg protein as illustrated by the anti-BMP-4 immunoblot (inset and data not shown).

this interaction was specific as it could be competed by BMP-2 but not by a 10-fold excess of platelet-derived growth factor, epidermal growth factor, Activin or transforming growth factor (TGF)- β 1 (Fig. 4b). We used chemical crosslinking with disuccinimidyl suberate (DSS) to determine whether this molecular interaction was direct. After separation of the complexes under reducing conditions, haemagglutinin (HA)-tagged xTsg (1 nM) formed mostly dimers and a small amount of monomers (Fig. 4c, lane 1). In the presence of 1 nM BMP-4, the xTsg dimers shifted to a molecular mass consistent with the binding of one BMP-4 dimer (Fig. 4c, lanes 2 and 5). Using an immunoprecipitation assay, the apparent dissociation constant for equilibrium binding of BMP-4 to xTsg (K_d) was determined by Scatchard analysis and found to be about 2.5 nM for both *Xenopus* and *Drosophila* Tsg (Fig. 4d and data not shown). To map the BMP binding domain, we generated constructs consisting of the conserved domains of xTsg and dTsg separated at the sites indicated in Fig. 1a and prepared secreted proteins. As shown in Fig. 5a and b, the BMP-binding activity resided in the N-terminal region of xTsg and dTsg, that is, in the domain that shares sequence similarities with the BMP-binding modules of Chordin. We conclude that xTsg is a secreted BMP-binding protein that functions in embryos as an agonist of BMP signalling. The other secreted BMP-binding proteins identified, such as Chordin, Noggin, Follistatin and members of the Cerberus family, are antagonists of BMP activity^{2,32–34}.

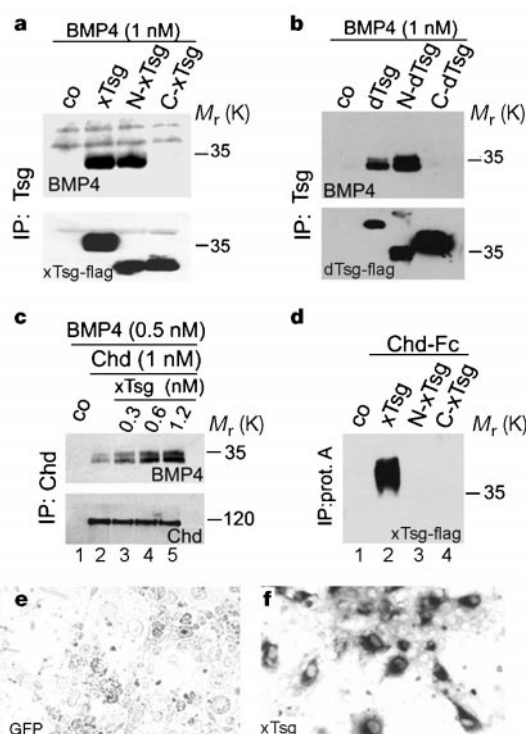


Figure 5 The Tsg N-terminal domain is sufficient to interact with BMP-4, but not with Chordin. **a,b**, Western blot analyses of BMP-4 (1 nM) bound to 1 nM of *Xenopus* or *Drosophila* Tsg. As a loading control, the same membranes were stripped and probed against Tsg (bottom). **c**, Western blot analyses of BMP-4 (0.5 nM) bound by 1 nM Chordin (lane 2) in the presence of 0.3, 0.6 and 1.2 nM xTsg (lanes 3, 4 and 5, respectively). The presence of xTsg increases the binding of BMP-4 to Chordin about fourfold (quantification by Phosphorimager, not shown). Bottom, same membrane probed with anti-Chordin antibody. **d**, Immunoprecipitation with protein A-sepharose of a mixture of 1 nM Chordin–Fc fusion protein with 1 nM xTsg, N-xTsg and C-xTsg. **e,f**, The Chordin–xTsg interaction was confirmed by binding to COS cells transfected with green fluorescent protein (negative control) or with xTsg cDNA, which were permeabilized and stained with a Chordin–alkaline phosphatase fusion protein.

xTsg competes with CR-1 to bind BMP

We next tested whether Chordin and xTsg could compete for BMP binding. Full-length Chordin (1 nM) was immunoprecipitated in the presence of 0.5 nM BMP-4 and increasing amounts of xTsg protein. Instead of competing for the limited amounts of BMP-4, xTsg stimulated binding of BMP-4 to full-length Chordin (Fig. 5c, lanes 2–5). Further analyses showed that xTsg itself could bind to Chordin even in the absence of added BMP (Fig. 5d, lane 2). This binding was confirmed using a cell-culture assay in which permeabilized cells transfected with xTsg cDNA were stained with a Chordin–alkaline phosphatase fusion protein^{35,36} (Fig. 5e, f). The binding of xTsg to Chordin required intact xTsg, as neither the N- nor the C-terminal fragments sufficed for the interaction (Fig. 5d, lanes 3 and 4). The observation that the N terminus is unable to bind Chordin but can still bind BMP-4 suggests that these two interactions occur through different sites.

Crosslinking experiments with DSS showed that xTsg stimulates the binding of BMP-4 to full-length Chordin by forming a trimolecular complex with a relative molecular mass of about 220,000 (M_r , 220K) (Fig. 6a, lane 3). This shift is consistent with the crosslinking of one dimer of BMP-4 and one dimer of xTsg per Chordin monomer, as depicted in Fig. 6b. To investigate the molecular mechanism by which xTsg can antagonize the activity of CR1, equimolar amounts (1 nM) of BMP-4, CR1 and xTsg were incubated for 1 h at room temperature before crosslinking with DSS. When CR1 and BMP-4 were incubated together, a complex of M_r 50K was formed, corresponding to the binding of a CR1 monomer to a BMP dimer (Fig. 6a, compare lanes 1 and 4). Incubation of xTsg

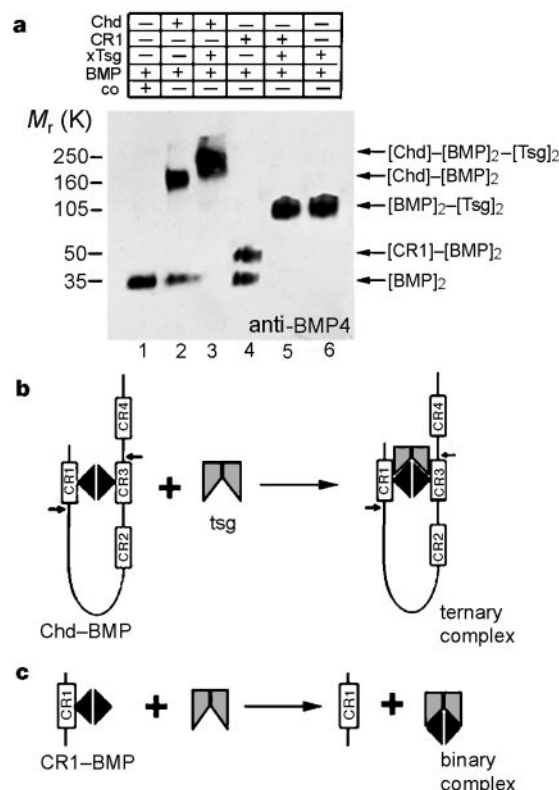


Figure 6 xTsg competes for binding of BMP-4 with CR1 but not with full-length Chordin. Crosslinking analyses were performed using DSS. **a**, Lane 1, BMP-4 (1 nM). 1 nM Chd and 1 nM BMP-4 form a complex of about 150K (lane 2) that shifts to ~220K in the presence of 1 nM xTsg (lane 3). When 1 nM CR1 was crosslinked to BMP-4 a 50K complex was detected (lane 4). When xTsg was added a complex of around 100K was produced exclusively (lane 5), which was identical to the $[BMP]_2$ – $[xTsg]_2$ complex (lane 6). **b,c**, Diagrams summarizing the crosslinking results. The small arrows in Chordin indicate the Xoloid cleavage sites.

and BMP-4 resulted in the formation of a complex of about 100K, corresponding to a dimer of xTsg bound to a dimer of BMP-4 (Fig. 6a, lane 6). Remarkably, when CR1, BMP-4 and xTsg were incubated together, only the [xTsg]₂–[BMP]₂ complex was formed (Fig. 6a, lane 5). Furthermore, when BMP-4 and CR1 were preincubated for 1 h, BMP-4 could still be dislodged from these preformed complexes by addition of xTsg (not shown). We conclude from these results that, in contrast to full-length Chordin, CR1 bound to BMP-4 is released in the presence of xTsg, resulting in a binary complex of BMP-4 and xTsg (Fig. 6c).

Endogenous xTsg antagonizes CR1 activity

To investigate the effect of loss of function of endogenous xTsg, we used two complementary approaches. A construct secreting the C-terminal conserved domain (*dn-xTsg*) was found to have dominant-negative effects (see below) and double-stranded xTsg RNA (RNAi) was also found to interfere with endogenous xTsg function. RNAi potently and specifically inhibits gene expression in a number of species^{37–39}. xTsg RNAi (100 pg) was injected into each animal blastomere at the eight-cell stage together with epitope-tagged xTsg and CR1 mRNAs. As shown in the inset of Fig. 7h, RNAi inhibited the expression of secreted xTsg but not of CR1 protein, indicating that RNAi is effective and specific in *Xenopus*. Microinjection of *dn-xTsg* or of RNAi resulted in abnormal development of the postanal region, in particular loss of the ventral fin and shortening of the tail which were only evident at swimming tadpole stages (data not shown). At earlier stages of development, only mild defects in the perianal region were observed (Fig. 7d, e); these minimal early phenotypes may be due to the strong maternal contribution of xTsg (Fig. 1b). However, both *dn-xTsg* mRNA and RNAi greatly potentiated the dorsalizing effects of low doses of CR1 mRNA injected into all blastomeres of four-cell embryos (compare Fig. 7a, f and h). These effects could be partially rescued by co-injection of wild-type xTsg mRNA, although the rescue was more effective for *dn-xTsg* (Fig. 7f, g) than for xTsg RNAi (Fig. 7h, i).

The results could be confirmed using molecular markers in ectodermal explants. As can be seen in Fig. 7j, *dn-xTsg* and xTsg RNAi potentiated the induction of the neural marker NCAM, as well as the inhibition of the epidermal marker *cytokeratin*, by CR1 mRNA (lanes 6, 8 and 10). The neuralizing effect of CR1, but not that of *tBR*, could be reversed by xTsg mRNA (lanes 6 and 7, and data not shown). The strong neuralizing effects observed in the presence of *dn-xTsg* or RNAi could also be reversed by full-length xTsg mRNA (lanes 9 and 11). These loss-of-function experiments indicate that endogenous levels of xTsg activity are sufficient to inhibit the dorsalizing effects of microinjected CR1.

Discussion

The *Xenopus* homologue of *dTsg* is a member of the BMP-4 synexpression group²⁶, and is transcribed in the ventral-most region of the embryo. Overexpression of this secreted protein ventralizes the embryo through a molecular mechanism functioning upstream of the BMP receptor (Fig. 3b). Epistasis experiments support a model in which xTsg would promote BMP signalling downstream of Chordin cleavage by the metalloproteinase Xolloid. Chordin is abundantly secreted by the dorsal pole of the embryo, reaching concentrations of 6–12 nM in the extracellular space of Spemann's organizer⁴⁰. Chordin binds BMPs through cysteine-rich modules¹². Xolloid cleaves Chordin downstream of the two cysteine-rich domains that have the highest affinity for BMP binding^{3,11,12}. These proteolytic digestion products retain residual BMP-binding activity and can still inhibit the interaction between BMP and its cognate receptor¹². Our results show that xTsg binds BMP directly in the low nanomolar range and will compete effectively with a cysteine-rich module for binding to BMP, leading to the formation of xTsg–BMP complexes that permit BMP signalling (Figs 2, 3 and 6). In this way, xTsg would promote BMP activity

by dislodging the growth factor from an inhibitor in the extracellular space. Interfering with endogenous xTsg by two independent approaches, *dn-xTsg* and RNAi, greatly increased the ability of CR1 to inhibit BMP. xTsg mRNA did not cooperate with BMP-4 mRNA in the induction of the target gene *Xvent-1* (ref. 26) in animal cap explants, nor did xTsg facilitate the binding of BMP-4 to recombinant BMP receptor IA (ref. 12) in biochemical assays (data not shown). This indicates that in the absence of Chordin cleavage

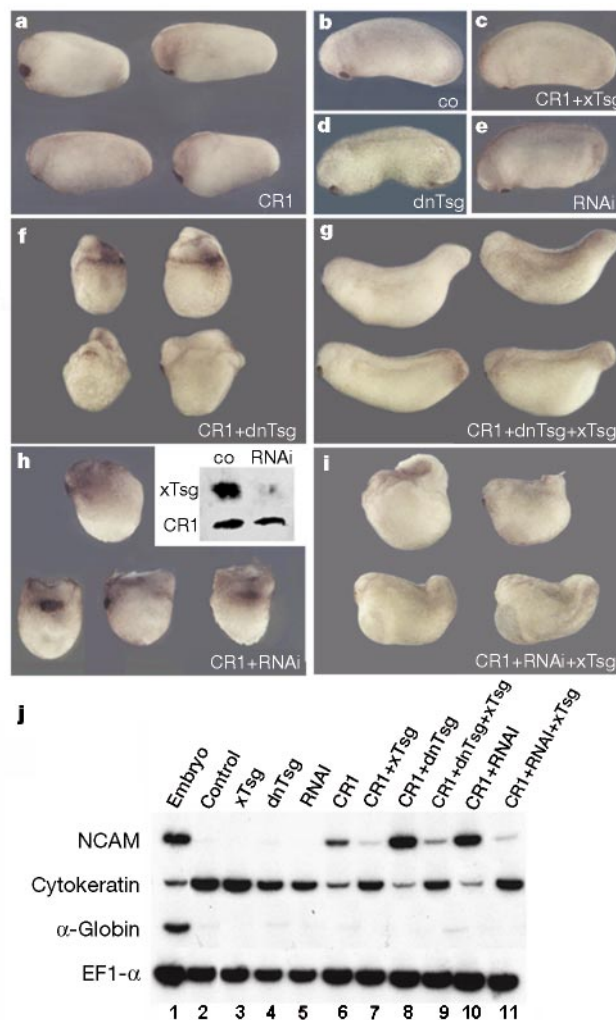


Figure 7 Loss of function of endogenous xTsg potentiates the activity of CR1 mRNA. Four-cell embryos were injected into each cell in the marginal region and cultured to stage 25. **a**, Injection of CR1 results in slight dorsalization compared with control uninjected embryos (**b**, **c**). Co-injection of xTsg and CR1 mRNA prevents dorsalization by CR1. Injection of *dn-xTsg* (**d**) or RNAi (**e**) disrupts the development of the perianal region leading to shortening of the tail at later stages (not shown); at these early tadpole stages the phenotype was minimal. Co-injection of CR1 and *dn-xTsg* mRNA (**f**) or RNAi (**h**) leads to strong synergistic dorsalization of the embryos, which can be partially rescued by co-injecting full-length xTsg mRNA (**g**, **i**). **h**, Inset, co-injection of HA-tagged xTsg mRNA and Myc-tagged CR1 mRNA with or without xTsg RNAi. Ectodermal explants were dissociated², incubated for 14 h at 16 °C, and secreted proteins analysed by western blot. **j**, RT-PCR analysis of stage 25 ectodermal explants of embryos injected at the eight-cell stage. NCAM induction reflects inhibition of BMP signalling by CR1 (lane 6), which is potentiated by co-injection of *dn-xTsg* or RNAi (lanes 8 and 10). The induction of NCAM by CR1, as well as its potentiation by co-injection of *dn-xTsg* or RNAi, was inhibited by co-injection of wild-type xTsg (lanes 7, 9, 11). The epidermal marker gene *cytokeratin* is regulated reciprocally by the injected RNAs and serves as a reporter of endogenous BMP signals. α-Globin was used as a mesodermal marker and EF1-α as loading control. The amounts of mRNA per blastomere injected were CR1, 20 pg; xTsg, 200 pg; *dn-xTsg*, 100 pg; and RNAi, 100 pg.

products xTsg does not increase BMP signalling. Unlike CR1, the binding of full-length Chordin to BMP-4 is enhanced by xTsg, leading to the formation of a ternary complex that can be cross-linked *in vitro*; perhaps xTsg results in more efficient cleavage of Chordin by Xolloid.

The initial impetus to investigate whether xTsg bound BMP was provided by the observation that the N-terminal conserved domain shared amino-acid similarities with Chordin cysteine-rich repeats. This similarity was found to be functionally significant, as this domain contains the BMP-binding activity of xTsg. Cysteine-rich modules of the Chordin type are found in many extracellular proteins such as von Willebrand factor, thrombospondin, Nel and fibrillar procollagens. The cysteine-rich domain in procollagen IIA binds BMP and TGF- β 1 (refs 12, 41) and is responsible for the dorsalizing activity of procollagen IIA in *Xenopus*¹². It is therefore possible that the molecular mechanism described here for xTsg may provide a more general paradigm for signalling in the extracellular space. Similar functions in the release of latent growth factors bound to cysteine-rich domains could be executed by other secreted proteins that share structural similarities with the N-terminal repeat of xTsg, such as members of the connective tissue growth factor (CTGF) and insulin-like growth factor-binding protein (IGFBP) families^{24,42}. In addition, the Cripto and Oep (one-eyed pinhead) proteins, which provide a permissive function for signalling by nodal TGF- β family members in mouse and zebrafish^{43,44}, share sequence similarities to xTsg in their conserved EGF-like domains (data not shown).

In *Drosophila*, loss of function of dTsg results in the loss of amnioserosa, which requires maximal amounts of Dpp/Scw signalling^{14–16}, but does not affect the rest of the dorsal–ventral pattern²⁴. Overexpression of dTsg driven by a promoter expressed in the ventral-most region of the embryo rescues dorsal amnioserosa formation in a dTsg mutant background, indicating that the protein can diffuse throughout the embryo²⁵. As dTsg does not affect other dorsal–ventral fates in *Drosophila*, it was thought to provide a permissive signal required for amnioserosa differentiation after cells have been exposed to peak levels of Dpp/Scw signalling^{24,25}. Our results show that *Drosophila* Tsg can bind BMPs through its N-terminal domain (Fig. 5b). *Drosophila* Sog, although not yet demonstrated to bind Dpp/Scw/BMP in biochemical studies, is cleaved by Tolloid at three sites in the presence of BMP⁴⁵. Two of these cleavage sites are located at similar positions to those in *Xenopus* Chordin³. We propose that dTsg can act only in the amnioserosa because its function is to release active Dpp/Scw from the cleavage fragments of Sog by Tolloid. Peak signalling would be mediated by Dpp/Scw originating in more ventral regions and reaching the dorsal region by diffusion as a complex with Sog²¹. Release would occur only in dorsal-most regions in which maximal levels of dTsg, Tolloid, and fragments of Sog bound to BMP are prevalent. In more ventral regions, full-length free Sog would be in excess²¹ and would inhibit any released BMPs. This would be particularly true in the presence of dTsg in dorso-lateral regions, in view of the observation that xTsg increases the binding of BMP to full-length Chordin (Fig. 5c). This model has the attraction of explaining the paradoxical effects of Sog, which inhibits Dpp/Scw signalling in ventral ectoderm but promotes it in the most dorsal regions of the embryo^{20,46}. By introducing dTsg as an additional factor that permits the release of Dpp/Scw from inactive complexes in the amnioserosa, it is not necessary to invoke positive (that is, non-inhibitory) signalling effects of Sog proteolytic cleavage products to account for the Tolloid-dependent long-range effects of Sog diffusion on peak Dpp signalling^{21–23}.

One of the great surprises of developmental genetics has been the evolutionary conservation of molecular mechanisms. The best known example is the discovery that conserved sets of Hox genes determine the anterior–posterior axis in all bilateral animals^{47,48}. The activities and expression patterns of Dpp/BMP-4, Sog/Chordin

and Tolloid/Xolloid have led to the view that the dorsal–ventral axis has been inverted in the course of evolution^{1,18}. We now show that *Drosophila* Tsg, a gene expressed in the dorsal-most blastoderm, has a vertebrate homologue expressed in the ventral-most tissues of the *Xenopus* embryo. xTsg is part of an intricate extracellular signalling pathway in which the two poles of the dorsal–ventral pattern are defined by sources of secreted Sog/Chordin and dTsg/xTsg in opposite sides of the embryo. The results are consistent with the idea that the ventral side of the arthropod is homologous to the dorsal side of the vertebrate, as proposed by Geoffroy Saint-Hilaire⁴⁹.

Note added in proof: As this work was being reviewed, it was reported⁵⁰ that alternative Sog processing may generate protein forms with increased BMP inhibitory activity. In the light of our data, some of those findings might be explained by the formation of trimolecular complexes consisting of Tsg, Dpp and Sog fragments that are insensitive to Tolloid cleavage. □

Methods

DNA constructs

We screened a *Xenopus* gastrula cDNA library with the coding sequence of the hTsg EST; a full-length clone was isolated, sequenced and subcloned in the expression vector pCS2⁺. xTsg and dTsg were epitope-tagged by PCR either using a 3' primer including an HA tag before the stop codon or by amplification of xTsg or dTsg subclones lacking the leader sequence and cloning into an expression vector including the chordin signal peptide and a Flag tag sequence (pChd–Flag, constructed by S. Piccolo). pChd–Flag was used to generate N-xTsg, C-xTsg, N-dTsg and C-dTsg. The *Xenopus* pCS2–CR1 construct was generated by PCR from pCS2–chordin³ using a 5' vector primer and a 3' primer including Chordin sequences up to Asn 146, followed by a Myc tag.

Protein binding and crosslinking

Proteins were obtained by transient transfection of 293T cells (xTsg, dTsg) or S2 cells (Ch–dAP, Chd–Fc). Immunoprecipitations and equilibrium binding assays were performed as described¹². For crosslinking⁴⁰, we incubated protein mixtures containing BMP-4 (R&D Systems) for 1 h at room temperature in 80 μ l of PBS; disuccinimidyl suberate (Pierce) was then added to a final concentration of 0.85 mM and incubated at room temperature for a further 30 min. The reaction was stopped by adding 4 μ l 1 M Tris-HCl (pH 7.4) and samples were electrophoresed in SDS gels under reducing conditions. For secretion-trap cell staining³⁵, COS-7 cells were transiently transfected and after two days fixed (1.8% paraformaldehyde for 15 min), permeabilized (0.1% TritonX-100 for 15 min), and incubated for 1 h at room temperature with Chordin–alkaline phosphatase fusion protein. Embryo manipulations and RNA analyses were as described¹².

Received 20 March; accepted 23 May 2000.

- De Robertis, E. M. & Sasai, Y. A common plan for dorsoventral patterning in Bilateria. *Nature* **380**, 37–40 (1996).
- Harland, R. & Gerhart, J. Formation and function of Spemann's organizer. *Annu. Rev. Cell. Dev. Biol.* **13**, 611–667 (1997).
- Piccolo, S. *et al.* Cleavage of Chordin by Xolloid metalloprotease suggests a role for proteolytic processing in the regulation of Spemann organizer activity. *Cell* **91**, 407–416 (1997).
- Weinmaster, G. Reprolysins and astacins. *Science* **279**, 336–337 (1998).
- Schulte-Merker, S., Lee, K. J., McMahon, A. P. & Hammerschmidt, M. The zebrafish organizer requires chordin. *Nature* **387**, 862–863 (1997).
- Kishimoto, Y., Lee, K. H., Zon, L., Hammerschmidt, M. & Schulte-Merker, S. The molecular nature of zebrafish swirl: BMP2 function is essential during early dorsoventral patterning. *Development* **124**, 4457–4466 (1997).
- Dick, A. *et al.* Essential role of Bmp7 (snailhouse) and its prodomain in dorsoventral patterning of the zebrafish embryo. *Development* **127**, 343–354 (2000).
- Schmid, B. *et al.* Equivalent genetic roles for bmp7/snailhouse and bmp2b/swirl in dorsoventral pattern formation. *Development* **127**, 957–967 (2000).
- Hild, M. *et al.* The smad5 mutation somitabun blocks Bmp2b signaling during early dorsoventral patterning of the zebrafish embryo. *Development* **126**, 2149–2159 (1999).
- Connors, S. A., Trout, J., Ekker, M. & Mullins, M. C. The role of tolloid/mini fin in dorsoventral pattern formation of the zebrafish embryo. *Development* **126**, 3119–3130 (1999).
- Scott, I. C. *et al.* Mammalian BMP-1/Tolloid-related metalloproteinases, including novel family member mammalian Tolloid-like 2, have differential enzymatic activities and distributions of expression relevant to patterning and skeletogenesis. *Dev. Biol.* **213**, 283–300 (1999).
- Larrain, J. *et al.* BMP-binding modules in Chordin: a model for signalling regulation in the extracellular space. *Development* **127**, 821–830 (2000).
- Ferguson, E. L. & Anderson, K. V. Localized enhancement and repression of the activity of the TGF- β family member, decapentaplegic, is necessary for dorsal–ventral pattern formation in the *Drosophila* embryo. *Development* **114**, 583–597 (1992).
- Ferguson, E. L. & Anderson, K. V. Decapentaplegic acts as a morphogen to organize dorsal–ventral pattern in the *Drosophila* embryo. *Cell* **71**, 451–461 (1992).
- Neul, J. L. & Ferguson, E. L. Spatially restricted activation of the SAX receptor by SCW modulates DPP/TKV signaling in *Drosophila* dorsal–ventral patterning. *Cell* **95**, 483–494 (1998).

16. Nguyen, M., Park, S., Marques, G. & Arora, K. Interpretation of a BMP activity gradient in *Drosophila* embryos depends on synergistic signaling by two type I receptors, SAX and TKV. *Cell* **95**, 495–506 (1998).
17. François, V., Solloway, M., O'Neill, J. W., Emery, J. & Bier, E. Dorsal–ventral patterning of the *Drosophila* embryo depends on a putative negative growth factor encoded by the short gastrulation gene. *Genes Dev.* **8**, 2602–2616 (1994).
18. Holley, S. A. *et al.* A conserved system for dorsal–ventral patterning in insects and vertebrates involving sog and chordin. *Nature* **376**, 249–253 (1995).
19. Schmidt, J., François, V., Bier, E. & Kimelman, D. *Drosophila* short gastrulation induces an ectopic axis in *Xenopus*: evidence for conserved mechanisms of dorsal–ventral patterning. *Development* **121**, 4319–4328 (1995).
20. Zusman, S. B., Sweeton, D. & Wieschaus, E. F. Short gastrulation, a mutation causing delays in stage-specific cell shape changes during gastrulation in *Drosophila melanogaster*. *Dev. Biol.* **129**, 417–427 (1988).
21. Holley, S. A. *et al.* The *Xenopus* dorsalizing factor noggin ventralizes *Drosophila* embryos by preventing DPP from activating its receptor. *Cell* **86**, 607–617 (1996).
22. Bier, E. A unity of opposites. *Nature* **398**, 375–376 (1999).
23. Ashe, H. L. & Levine, M. Local inhibition and long-range enhancement of Dpp signal transduction by Sog. *Nature* **398**, 427–431 (1999).
24. Mason, E. D., Konrad, K. D., Webb, C. D. & Marsh, J. L. Dorsal midline fate in *Drosophila* embryos requires twisted gastrulation, a gene encoding a secreted protein related to human connective tissue growth factor. *Genes Dev.* **8**, 1489–1501 (1994).
25. Mason, E. D., Williams, S., Grotendorst, G. R. & Marsh, J. L. Combinatorial signaling by Twisted gastrulation and Decapentaplegic. *Mech. Dev.* **64**, 61–75 (1997).
26. Niehrs, C. & Pollet, N. Synexpression groups in eukaryotes. *Nature* **402**, 483–487 (1999).
27. Fainsod, A., Steinbeisser, H. & De Robertis, E. M. On the function of BMP-4 in patterning the marginal zone of the *Xenopus* embryo. *EMBO J.* **13**, 5015–5025 (1994).
28. Onichtchouk, D. *et al.* Silencing of TGF- β signalling by the pseudoreceptor BAMBI. *Nature* **401**, 480–485 (1999).
29. Green, J. B. & Smith, J. C. Graded changes in dose of a *Xenopus* activin A homologue elicit stepwise transitions in embryonic cell fate. *Nature* **347**, 391–394 (1990).
30. Goodman, S. A. *et al.* BMP1-related metalloproteinases promote the development of ventral mesoderm in early *Xenopus* embryos. *Dev. Biol.* **195**, 144–157 (1998).
31. Jones, C. M., Lyons, K. M., Lapan, P. M., Wright, C. V. E. & Hogan, B. L. M. DVR-4 (Bone Morphogenetic Protein-4) as a posterior-ventralizing factor in *Xenopus* mesoderm induction. *Development* **115**, 639–647 (1992).
32. Hsu, D. R., Economides, A. N., Wang, X., Eimon, P. M. & Harland, R. M. The *Xenopus* dorsalizing factor Gremlin identifies a novel family of secreted proteins that antagonize BMP activities. *Mol. Cell* **1**, 673–683 (1998).
33. Piccolo, S. *et al.* The head inducer Cerberus is a multifunctional antagonist of Nodal, BMP and Wnt signals. *Nature* **397**, 707–710 (1999).
34. Rodríguez Esteban, C. *et al.* The novel Cer-like protein Caronte mediates the establishment of embryonic left-right asymmetry. *Nature* **401**, 243–251 (1999).
35. Davis, S. *et al.* Isolation of angiopoietin-1, a ligand for the TIE2 receptor, by secretion-trap expression cloning. *Cell* **87**, 1161–1169 (1996).
36. Flanagan, J. G. & Leder, P. The kit ligand: a cell surface molecule altered in steel mutant fibroblasts. *Cell* **63**, 185–194 (1990).
37. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
38. Tuschl, T., Zamore, P. D., Lehmann, R., Bartel, D. P. & Sharp, P. A. Targeted mRNA degradation by double-stranded RNA in vitro. *Genes Dev.* **13**, 3191–3197 (1999).
39. Boshier, J. M. & Labouesse, M. RNA interference: genetic wand and genetic watchdog. *Nature Cell Biol.* **2**, 31–36 (2000).
40. Piccolo, S., Sasai, Y., Lu, B. & De Robertis, E. M. Dorsal–ventral patterning in *Xenopus*: inhibition of ventral signals by direct binding of chordin to BMP-4. *Cell* **86**, 589–598 (1996).
41. Zhu, Y., Oganessian, A., Keene, D. R. & Sandell, L. J. Type IIA procollagen containing the cysteine-rich amino propeptide is deposited in the extracellular matrix of prechondrogenic tissue and binds to TGF- β 1 and BMP-2. *J. Cell Biol.* **144**, 1069–1080 (1999).
42. Grotendorst, G. R. Connective tissue growth factor: a mediator of TGF- β action on fibroblasts. *Cytokine Growth Factor Rev.* **8**, 171–179 (1997).
43. Zhang, J., Talbot, W. S. & Schier, A. F. Positional cloning identifies zebrafish one-eyed pinhead as a permissive EGF-related ligand required during gastrulation. *Cell* **92**, 241–251 (1998).
44. Schier, A. F. & Shen, M. M. Nodal signalling in vertebrate development. *Nature* **403**, 385–389 (2000).
45. Marques, G. *et al.* Production of a DPP activity gradient in the early *Drosophila* embryo through the opposing actions of the SOG and TLD proteins. *Cell* **91**, 417–426 (1997).
46. Biehs, B., François, V. & Bier, E. The *Drosophila* short gastrulation gene prevents Dpp from autoactivating and suppressing neurogenesis in the neuroectoderm. *Genes Dev.* **10**, 2922–2934 (1996).
47. De Robertis, E. M. in *Guidebook of Homeobox Genes* (ed. Duboule, D.) 11–23 (IRL, Oxford, 1994).
48. Knoll, A. H. & Carroll, S. B. Early animal evolution: emerging views from comparative biology and geology. *Science* **284**, 2129–2137 (1999).
49. Geoffroy Saint-Hilaire, E. Considérations générales sur la vertèbre. *Mém. Mus. Hist. Nat.* **9**, 89–119 (1822).
50. Yu *et al.* Processing of the *Drosophila* Sog protein creates a novel BMP inhibitory activity. *Development* **127**, 2143–2154 (2000).

Acknowledgements

We thank J. Fessler, L. Zipursky and members of our laboratory for comments on the manuscript, and U. Tran and A. Cuellar for technical assistance. M.O. and J.L. are HFSP and Pew postdoctoral fellows, respectively. This work was supported by the NIH and the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to E.M.D.R. (e-mail: derobert@hhmi.ucla.edu). The accession number for *xTsg* is AF245221. Human and mouse *Tsg* homologues are encoded by ESTs AA486291 and AW258143.

Kondo effect in an integer-spin quantum dot

S. Sasaki*, S. De Franceschi†, J. M. Elzerman†, W. G. van der Wiel†, M. Eto‡, S. Tarucha§ & L. P. Kouwenhoven†

* NTT Basic Research Laboratories, Atsugi-shi, Kanagawa 243-0129, Japan

† Department of Applied Physics, DIMES, and ERATO Mesoscopic Correlation Project, Delft University of Technology, PO Box 5046, 2600 GA Delft, The Netherlands

‡ Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan

§ ERATO Mesoscopic Correlation Project, University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan

The Kondo effect—a many-body phenomenon in condensed-matter physics involving the interaction between a localized spin and free electrons—was discovered in metals containing small amounts of magnetic impurities, although it is now recognized to be of fundamental importance in a wide class of correlated electron systems^{1,2}. In fabricated structures, the control of single, localized spins is of technological relevance for nanoscale electronics^{3,4}. Experiments have already demonstrated artificial realizations of isolated magnetic impurities at metallic surfaces^{5,6}, nanoscale magnets⁷, controlled transitions between two-electron singlet and triplet states⁸, and a tunable Kondo effect in semiconductor quantum dots^{9–12}. Here we report an unexpected Kondo effect in a few-electron quantum dot containing singlet and triplet spin states, whose energy difference can be tuned with a magnetic field. We observe the effect for an even number of electrons, when the singlet and triplet states are degenerate. The characteristic energy scale is much larger than in the ordinary spin-1/2 case.

Quantum dots are small electronic devices¹³, which confine a well-defined number of electrons, N . The total spin is zero or an integer for even N , and half-integer for odd N . The latter case is the canonical example of the Kondo effect^{14,15} when all electrons can be ignored, except for the one with the highest energy; that is, the case of a single, isolated spin, $S = 1/2$ (Fig. 1a). Although the energy level of the electron ϵ_0 is well below the Fermi energies of the two leads, Heisenberg uncertainty allows the electron on the dot to tunnel to one of the leads whereupon it is replaced quickly by another electron. The timescale for such a co-tunnelling process¹⁶ is $\sim \hbar/U$, where $\hbar = 2\pi\hbar$ is Planck's constant and U is the on-site Coulomb energy. Figure 1a illustrates that particle exchange by co-tunnelling can effectively flip the spin on the dot. At low temperature, the coherent superposition of all possible co-tunnelling processes involving spin flip can result in a time-averaged spin equal to zero. The whole system—that is, quantum dot plus electrodes—forms a spin singlet. The energy scale for this singlet state is the Kondo temperature, T_K . In terms of density of states, a narrow peak with a width $\sim k_B T_K$ develops at the Fermi energy (k_B is Boltzmann's constant). Note that for even N and $S = 0$, co-tunnelling gives rise to a lifetime broadening of the confined state, without producing any Kondo resonance. Such even/odd behaviour corresponding to no-Kondo/Kondo has been observed in recent experiments^{9,10}.

It is also possible that a quantum dot with even N has a total spin $S = 1$; for example, when the last two electrons have parallel spins. If the remaining $N - 2$ electrons can be ignored, this corresponds to a triplet state. Parallel spin filling is a consequence of Hund's rule occurring when the gain in exchange energy exceeds the spacing between single-particle states⁸. The spin of the triplet state can also be screened by co-tunnelling events. These are illustrated in the centre-left side of Fig. 1b. In contrast to single-particle states that are

considered in the spin-1/2 Kondo problem, the spin triplet consists of three degenerate two-particle states. Co-tunnelling exchanges only one of the two electrons with an electron from the leads. The total spin of the many-body Kondo state depends on how many modes in the leads couple effectively to the dot^{17,18}. If there is only one mode, the screening is not complete and the whole system does not reach a singlet state. In this case the Kondo effect is called “underscreened”. Calculations show that also for $S = 1$ a narrow Kondo resonance arises at the Fermi energy, but the corresponding T_K is typically lower than in the case of $S = \frac{1}{2}$ (refs 19, 20). Some experiments have reported the absence of even/odd behaviour^{21,22}, which may be related to the formation of higher spin states.

Here we investigate a quantum dot with even N , where the last two electrons occupy a degenerate state consisting of a spin singlet and a spin triplet. Figure 1b illustrates the different co-tunnelling processes occurring in this special circumstance. Starting from $|S = 1, S_z = 1\rangle$, where S_z is the z -component of the total spin on the dot, co-tunnelling via a virtual state $|1/2, 1/2\rangle$ can lead either to the triplet state $|1, 0\rangle$ or to the singlet state $|0, 0\rangle$. Via a second co-tunnelling event, the state $|1, -1\rangle$ can be reached. As for the $S = 1$ case, the local spin can fluctuate by co-tunnelling events. By

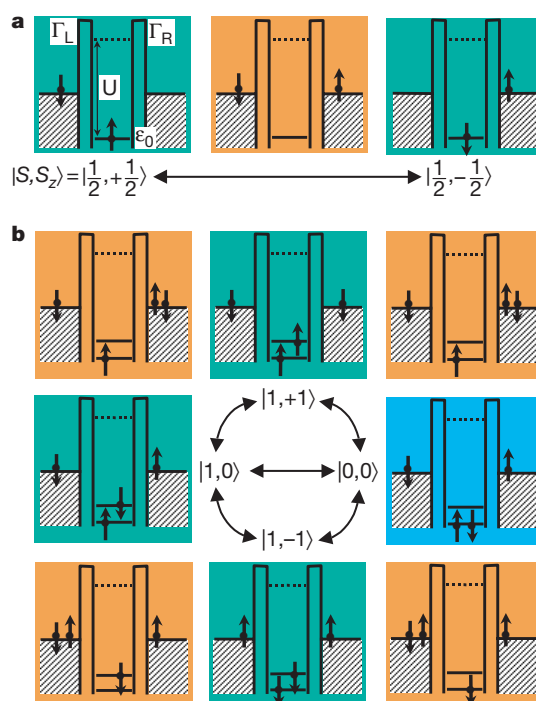


Figure 1 Spin-flip processes leading to ordinary and singlet-triplet Kondo effect in a quantum dot. **a**, Co-tunnelling event in a spin-1/2 quantum dot for odd N . Only the highest-energy electron is shown, occupying a single spin-degenerate level, ϵ_0 . (The case of two, or more, closely spaced levels has also been considered theoretically within the context of the spin-1/2 Kondo effect²⁰). The green panels refer to $S_z = 1/2$ and $-1/2$ ground states, which are coupled by a co-tunnelling event. The two tunnel barriers have tunnelling rates Γ_R and Γ_L . In the Coulomb blockade regime ($|\epsilon_0| \sim U$), adding or subtracting an electron from the dot implies an energy cost $\sim U$. Hence the intermediate step (orange panel) is a high-energy, virtual state. The spin-flip event depicted here is representative of a large number of higher-order processes which add up coherently such that the local spin is screened. This Kondo effect leads to an enhanced linear-response conductance at temperatures $T \lesssim T_K$. **b**, Co-tunnelling in an integer-spin quantum dot for even N at a singlet-triplet degeneracy. Two electrons can share the same orbital with opposite spins (singlet state in the blue panel) or occupy two distinct orbitals in one of the three spin-triplet configurations (green panels). The different spin states are coupled by virtual states (orange panels). As in the spin-1/2 case, spin-flip events can screen the local magnetic moment. Note that an $S = 1$ Kondo effect only involves $|1, +1\rangle$, $|1, 0\rangle$ and $|1, -1\rangle$.

coupling to all triplet states, the singlet state enhances the spin exchange interaction between the dot and the leads, resulting in a higher rate for spin fluctuations. This particular situation yields a strong Kondo effect, which is characterized by an enhanced T_K . This type of Kondo effect has not been considered before, probably because a singlet–triplet degeneracy does not occur in magnetic elements. Recent scaling calculations by Eto and Nazarov indeed indicate a strong enhancement of T_K at the singlet–triplet degeneracy²³; these workers also argue²³ that the total spin of the many-body Kondo state behaves as in the case of $S = 1$.

Our quantum dot has the external shape of a rectangular pillar (Fig. 2a, b) and an internal confinement potential close to a two-dimensional ellipse²⁴. The tunnel barriers between the quantum dot and the source and drain electrodes are thinner than in our previous devices^{8,24} such that co-tunnelling processes are enhanced. Figure 2d shows the linear response conductance (d.c. bias voltage $V_{sd} = 0$) versus gate voltage, V_g , and magnetic field, B . Regions shown dark blue have low conductance, and correspond to the regimes of Coulomb blockade for $N = 3$ to 10. In contrast to previous experiments^{9–12} on the Kondo effect, all performed on lateral

quantum dots with unknown electron number, here the number of confined electrons is precisely known. Red stripes in Fig. 2d represent Coulomb peaks as high as $\sim e^2/h$. The B -dependence of the first two lower stripes reflects the ground-state evolution for $N = 3$ and 4. Their similar B -evolution indicates that the third and fourth electron occupy the same orbital state with opposite spin, which is observed also for $N = 1$ and 2 (not shown). This is not the case for $N = 5$ and 6. The $N = 5$ state has $S = 1/2$, and the corresponding stripe shows a smooth evolution with B . Instead, the stripe for $N = 6$ has a kink at $B \approx 0.22$ T. From earlier analyses²⁴ and from measurements of the excitation spectrum at finite V_{sd} (discussed below), we can identify this kink with a transition in the ground state from a spin triplet to a spin singlet. At the triplet–singlet transition (at $B = B_0$ in Fig. 2c) we observe a strong enhancement of the conductance. In fact, over a narrow range around 0.22 T, the Coulomb gap for $N = 6$ disappears completely.

To explore this conductance anomaly, we show in Fig. 3a differential conductance measurements, dI/dV_{sd} versus V_{sd} , taken at $B = B_0$ and V_g corresponding to the dotted line in Fig. 2d. At temperature $T = 14$ mK, the narrow resonance around zero bias has a full-width at half-maximum, FWHM, of ~ 30 μ V. This is several times smaller than the lifetime broadening, $\Gamma = \Gamma_R + \Gamma_L \approx 150$ μ V, estimated from the FWHM of the Coulomb peaks. The height of the zero-bias resonance decreases logarithmically with T (Fig. 3b). These are typical ‘fingerprints’ of the Kondo effect. From $\text{FWHM} \approx k_B T_K$, we estimate $T_K \approx 350$ mK. We note that we can safely neglect the Zeeman spin splitting because $g\mu_B B_0 \approx 5$ μ V $\ll k_B T_K$, implying that the spin triplet is in fact three-fold degenerate at $B = B_0$. This condition is essential to the Kondo effect illustrated in Fig. 1b. Alternative schemes have recently been proposed for a Kondo effect where the degeneracy of the triplet state is lifted by a large magnetic field^{25,26}.

For $N = 6$, we find markedly anomalous T -dependence only at the

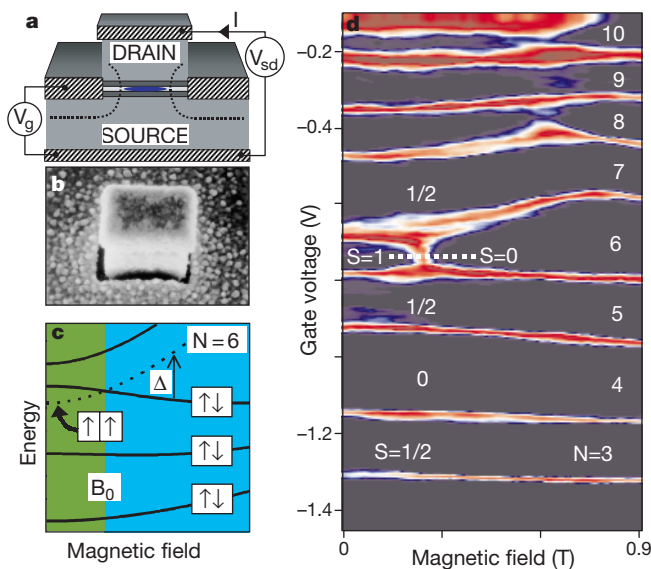


Figure 2 Sample description, energy spectrum and magnetic-field evolution of the ground state. **a**, Cross-section of our rectangular quantum dot. The semiconductor material consists of an undoped AlGaAs (7 nm)/InGaAs (12 nm)/AlGaAs (7 nm) double barrier structure sandwiched between n -doped GaAs source and drain electrodes. A gate electrode surrounds the pillar and is used to control the electrostatic confinement in the quantum dot. A d.c. bias voltage, V_{sd} , is applied between source and drain, and current, I , flows vertically through the pillar. In addition to V_{sd} , we apply a modulation with r.m.s. amplitude $V_{ac} = 3$ μ V at 17.7 Hz for lock-in detection. The gate voltage, V_g , can change the number of confined electrons, N , one-by-one from ~ 10 at $V_g = 0$ to 0 at $V_g = -1.8$ V. A magnetic field, B , is applied along the vertical axis. Temperature, T , is varied between 14 mK and 1 K. Our lowest effective electron temperature is 25 ± 5 mK. **b**, Scanning electron micrograph of a quantum dot with dimensions 0.45×0.6 μm^2 and height of ~ 0.5 μm . **c**, Schematic energy spectrum. Solid lines represent the B -evolution of the first four orbital levels in a single-particle model. The dashed line is obtained by subtracting the two-electron exchange coupling from the fourth level. At the crossing between this dashed line and the third orbital level at $B = B_0$ the ground state for $N = 6$ undergoes a triplet-to-singlet transition. $B_0 \approx 0.22$ T with a slight dependence on V_g . We define Δ as the energy difference between the triplet and the singlet states. **d**, Colour-scale representation of the linear conductance versus V_g and B . Red stripes denote conductance peaks of height $\sim e^2/h$. Blue regions of low conductance indicate Coulomb blockade. The V_g -position of the stripes reflects the ground-state evolution with B , for $N = 3$ to 10. The $N = 6$ ground state undergoes a triplet-to-singlet transition at $B_0 \approx 0.22$ T, which results in a conductance anomaly inside the corresponding Coulomb gap.

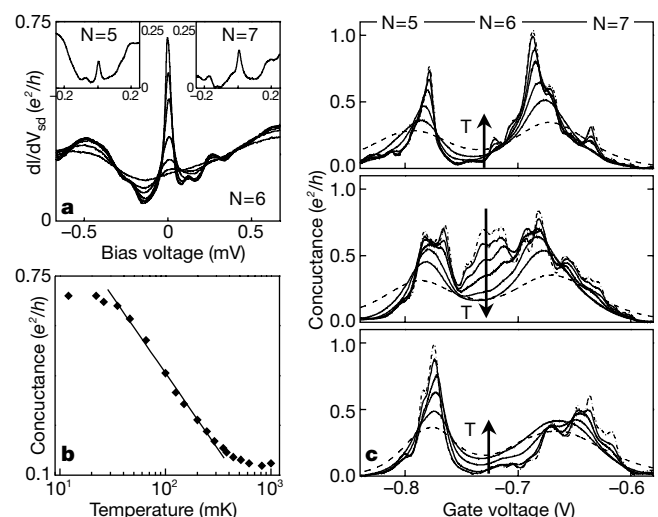


Figure 3 Zero-bias resonance and T -dependence of the conductance at the singlet–triplet degeneracy. **a**, Kondo resonance at the singlet–triplet transition. The dI/dV_{sd} versus V_{sd} curves are taken at $V_g = -0.72$ V, $B = 0.21$ T and for $T = 14, 65, 100, 200, 350, 520$ and 810 mK. Insets to **a**: Kondo resonances for $N = 5$ (left inset) and $N = 7$ (right inset), measured at $V_g = -0.835$ V and $V_g = -0.625$ V, respectively, and for $B = 0.11$ T and $T = 14$ mK. **b**, Peak height of zero-bias Kondo resonance versus T as obtained from **a** (filled diamonds). The line demonstrates a logarithmic T -dependence, which is characteristic of the Kondo effect. The saturation at low T is probably due to electronic noise. **c**, T -dependence of the linear conductance versus V_g for $B = 0.12$ T (spin-triplet ground state), $B = 0.22$ T (singlet–triplet degeneracy), and $B = 0.32$ T (spin-singlet ground state). Each panel shows 7 traces at $T = 20$ (dot-dashed line), 35, 70, 120, 260, 490 (solid lines) and 1,050 (dashed line) mK. The arrows emphasize the temperature dependence in the valley for $N = 6$.

singlet–triplet degeneracy. Figure 3c shows the conductance versus V_g for different T . The upper panel shows data for $B = 0.12$ T. The two Coulomb peaks correspond to the transition from $N = 5$ to 6 and from $N = 6$ to 7. The small, short-period modulations superimposed on the Coulomb peaks are due to a weak charging effect in the upper part of the GaAs pillar above the dot²⁷. We will ignore this fine structure and focus on the general T -dependence. Upon increasing T , the valley conductance for $N = 6$ goes up due to thermally activated transport. A similar behaviour is seen in the lower graph for $B = 0.32$ T. In contrast, at the singlet–triplet transition for $B = 0.22$ T we find an opposite T -dependence, again indicating the formation of a Kondo resonance. At the lowest T , the valley conductance is as high as $0.7e^2/h$, which is close to the height of the Coulomb peaks.

The T -dependence for $N = 5$ and 7 is visibly different from that in the non-Kondo valley for $N = 6$ (lower panel of Fig. 3c). Such a difference is a manifestation of the ordinary spin-1/2 Kondo effect expected for odd N . Indeed the corresponding zero-bias resonances are clearly observed (see insets to Fig. 3a). Their height, however, is much smaller than for the singlet–triplet Kondo effect. There is also some indication for a triplet Kondo effect in the T -dependence for $N = 6$ at $B = 0.12$ T, although the associated zero-bias anomaly is not as apparent.

We now investigate the effect of lifting the singlet–triplet degeneracy by changing B at a fixed V_g corresponding to the dotted line in Fig. 2d. Near the edges of this line, that is, away from B_0 , the Coulomb gap is well developed, as denoted by the dark colours in Fig. 2d. The dI/dV_{sd} versus V_{sd} traces still exhibit anomalies, however, now at finite V_{sd} (see Fig. 4a). For $B = 0.21$ T we observe

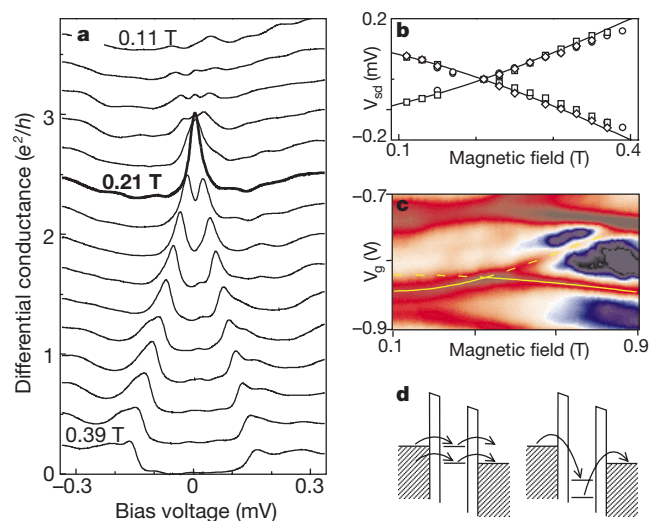


Figure 4 Singlet–triplet energy separation tuned by a magnetic field. **a**, dI/dV_{sd} versus V_{sd} characteristics taken along the dotted line in Fig. 2d ($V_g = -0.72$ V) at equally spaced magnetic fields $B = 0.11, 0.13, \dots, 0.39$ T. Curves are offset by $0.25 e^2/h$. **b**, Extracted peak positions from dI/dV_{sd} curves in **a** versus B . Each symbol refers to a gate voltage taken between -0.72 and -0.70 V. **c**, Colour-scale plot of dI/dV_{sd} measured at $V_{sd} = 0.67$ mV as a function of B and V_g . The solid line identifies the ground state, whereas the dashed line indicates the first excited state⁸ for $N = 6$. From their difference we extract the singlet–triplet energy splitting, $\Delta(B)$, using a proportionality factor $\alpha = 6.7 \text{ meV V}^{-1}$ to convert gate voltage into energy. The two solid lines in **b** represent $\pm\Delta(B)$ with a horizontal shift of 0.08 T to compensate for the shift of the singlet–triplet transition to a higher magnetic field in a high-bias measurement. **d**, Energy diagrams for two different transport regimes, both with $eV_{sd} = \Delta$. Left: both ground and excited states lie between the two Fermi energies, so two channels are available for direct tunnelling. The excitation spectrum in **c** is measured in this regime. Right: both ground and excited states lie below the Fermi levels of the leads (Coulomb blockade regime). Inelastic co-tunnelling is illustrated, where one electron tunnels out of the lower-energy state and another tunnels into the higher-energy state.

the singlet–triplet Kondo resonance at $V_{sd} = 0$. At higher B , this resonance splits apart showing two peaks at finite V_{sd} . It is important to note that these peaks occur inside the Coulomb gap. They result from “inelastic” co-tunnelling events^{16,28}, where “inelastic” refers to energy exchanging between quantum dot and electrodes (see right drawing in Fig. 4d). The upper traces in Fig. 4a, for $B < 0.21$ T, also show peak structures, although less pronounced.

In Fig. 4b we plot the positions of the dI/dV_{sd} peaks in the plane of V_{sd} and B . The different symbols refer to different gate voltages, and it can be seen that these positions do not depend on V_g . The solid lines are obtained from the excitation spectrum measured (see Fig. 4c) in direct tunnelling (see left-hand drawing in Fig. 4d). They represent the measured B -dependence of the singlet–triplet energy difference, Δ . The fact that these independent measurements coincide implies that inelastic co-tunnelling occurs when $eV_{sd} = \pm\Delta$. Note that this condition is independent of V_g , consistent with our observations. We believe that for small Δ ($\approx k_B T_K$), the split resonance reflects the singlet–triplet Kondo anomaly shifted to finite bias. This resembles the splitting of the Kondo resonance by the Zeeman effect^{9,10,29}, although on a very different B -scale. In the present case, the splitting occurs between two different multiparticle states and originates from the B -dependence of the orbital motion. For increasing Δ , the shift to larger V_{sd} induces spin-decoherence processes, which broaden and suppress the finite-bias peaks²⁹. For $B \approx 0.39$ T the peaks have evolved into steps²⁸, which may indicate that the spin-coherence associated with the Kondo effect has completely vanished. □

Received 3 February; accepted 28 April 2000.

- Hewson, A. C. *The Kondo Problem to Heavy Fermions* (Cambridge University Press, 1993).
- Cox, D. L. & Maple, M. B. Electronic pairing in exotic superconductors. *Phys. Today* **48**, 32–40 (1995).
- Prinz, G. A. Magnetoelectronics. *Science* **282**, 1660–1663 (1998).
- Loss, D. & DiVincenzo, D. P. Quantum computation with quantum dots. *Phys. Rev. A* **57**, 120–126 (1998).
- Madhavan, V., Chen, W., Jamneala, T., Crommie, M. F. & Wingreen, N. S. Tunneling into a single magnetic atom: spectroscopic evidence of the Kondo resonance. *Science* **280**, 567–569 (1998).
- Li, J., Schneider, W.-D., Berndt, R. & Delley, B. Kondo scattering observed at a single magnetic impurity. *Phys. Rev. Lett.* **80**, 2893–2896 (1998).
- Awschalom, D. D. & DiVincenzo, D. P. Complex dynamics of mesoscopic magnets. *Phys. Today* **48**, 43–48 (1995).
- Tarucha, S., Austing, D. G., Tokura, Y., van der Wiel, W. G. & Kouwenhoven, L. P. Direct Coulomb and exchange interaction in artificial atoms. *Phys. Rev. Lett.* **84**, 2485–2488 (2000).
- Goldhaber-Gordon, D. et al. Kondo effect in a single-electron transistor. *Nature* **391**, 156–159 (1998).
- Cronenwett, S. M., Oosterkamp, T. H. & Kouwenhoven, L. P. A tunable Kondo effect in quantum dots. *Science* **281**, 540–544 (1998).
- Schmid, J., Weis, J., Eberl, K. & von Klitzing, K. A quantum dot in the limit of strong coupling to reservoirs. *Physica B* **256–258**, 182–185 (1998).
- Simmel, E., Blick, R. H., Kotthaus, J. P., Wegscheider, W. & Bichler, M. Anomalous Kondo effect in a quantum dot at nonzero bias. *Phys. Rev. Lett.* **83**, 804–807 (1999).
- Kouwenhoven, L. P. et al. in *Mesoscopic Electron Transport* (eds Sohn, L. L., Kouwenhoven, L. P. & Schön, G.) 105–214 (NATO ASI Ser. E 345, Kluwer, Dordrecht, 1997).
- Glazman, L. I. & Raikh, M. E. Resonant Kondo transparency of a barrier with quasilocal impurity states. *JETP Lett.* **47**, 452–455 (1988).
- Ng, T. K. & Lee, P. A. On-site Coulomb repulsion and resonant tunneling. *Phys. Rev. Lett.* **61**, 1768–1771 (1988).
- Averin, D. V. & Nazarov, Y. V. in *Single Charge Tunneling* (eds Grabert, H. & Devoret, M. H.) 217–247 (Proc. NATO ASI Ser. B 294, Plenum, New York, 1991).
- Mattis, D. C. Symmetry of ground state in a dilute magnetic metal alloy. *Phys. Rev. Lett.* **19**, 1478–1481 (1967).
- Nozières, P. & Blandin, A. Kondo effect in real metals. *J. Phys.* **41**, 193–211 (1980).
- Wan, Y., Phillips, P. & Li, Q. Suppression of the Kondo effect in quantum dots by even-odd asymmetry. *Phys. Rev. B* **51**, 14782–14785 (1995).
- Izumida, W., Sakai, O. & Shimizu, Y. Kondo effect in single quantum dot systems—study with NRG method. *J. Phys. Soc. Jpn* **67**, 2444–2454 (1998).
- Maurer, S. M., Patel, S. R., Marcus, C. M., Duruöz, C. I. & Harris, J. S. Jr Coulomb blockade fluctuations in strongly coupled quantum dots. *Phys. Rev. Lett.* **83**, 1403–1406 (1999).
- Schmid, J., Weis, J., Eberl, K. & von Klitzing, K. Absence of odd-even parity behaviour for Kondo resonances in quantum dots. *Phys. Rev. Lett.* (submitted).
- Eto, M. & Nazarov, Y. V. Enhancement of Kondo effect in quantum dots with even number of electrons. Preprint cond-mat/0002010 (cited 29 May 2000) at (<http://xxx.lanl.gov>) (2000).
- Austing, D. G. et al. Ellipsoidal deformation of vertical quantum dots. *Phys. Rev. B* **60**, 11514–11523 (1999).
- Pustilnik, M., Avishai, Y. & Kikoin, K. Quantum dots with even number of electrons: Kondo effect in a finite magnetic field. *Phys. Rev. Lett.* **84**, 1756–1759 (2000).
- Giuliano, D. & Tagliacozzo, A. Spin fractionalization of an even number of electrons in a quantum dot. *Phys. Rev. Lett.* **84**, 4677–4680 (2000).
- Sivan, U. et al. Spectroscopy, electron-electron interaction, and levels in a disordered quantum dot. *Europhys. Lett.* **25**, 605–611 (1994).

28. Funabashi, Y., Ohtsubo, K., Eto, M. & Kawamura, K. Phase relaxation and non-equilibrium transport properties through multilevel quantum dot. *Jpn J. Appl. Phys.* **38**, 388–391 (1999).
29. Wingreen, N. S. & Meir, Y. Anderson model out of equilibrium: Noncrossing-approximation approach to transport through a quantum dot. *Phys. Rev. B* **49**, 11040–11052 (1994).
30. Inoshita, T., Shimizu, A., Kuramoto, Y. & Sakaki, H. Correlated electron transport through a quantum dot: The multiple-level effect. *Phys. Rev. B* **48**, 14725–14728 (1993).

Acknowledgements

We thank Yu. V. Nazarov, K. Majjala, S. M. Cronenwett, J. E. Mooij and Y. Tokura for discussions. We acknowledge financial support from the Specially Promoted Research, Grant-in-Aid for Scientific Research, from the Ministry of Education, Science and Culture in Japan, from the Dutch Organisation for Fundamental Research on Matter (FOM), from the NEDO joint research program, and from the EU via a TMR network.

Correspondence and requests for materials should be addressed to S.D.F. (e-mail: silvano@qt.tn.tudelft.nl).

Direct observation of the alignment of ferromagnetic spins by antiferromagnetic spins

F. Nolting*, A. Scholl*, J. Stöhr†, J. W. Seo‡§, J. Fompeyrine§, H. Siegart§, J.-P. Locquet§, S. Anders*, J. Lüning†, E. E. Fullerton†, M. F. Toney†, M. R. Scheinfein|| & H. A. Padmore*

* Advanced Light Source, 1 Cyclotron Road, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

† IBM Research Division, Almaden Research Center, 650 Harry Road, San Jose, California 95120, USA

‡ Institute de Physique, University of Neuchâtel, CH-2000 Neuchâtel, Switzerland

§ IBM Research Division, Zürich Research Laboratory, CH-8803 Rüschlikon, Switzerland

|| Department of Physics and Astronomy, Arizona State University, Tempe, Arizona 85287-1504, USA

The arrangement of spins at interfaces in a layered magnetic material often has an important effect on the properties of the material. One example of this is the directional coupling between the spins in an antiferromagnet and those in an adjacent ferromagnet, an effect first discovered¹ in 1956 and referred to as exchange bias. Because of its technological importance for the development of advanced devices such as magnetic read heads² and magnetic memory cells³, this phenomenon has received much attention^{4,5}. Despite extensive studies, however, exchange bias is still poorly understood, largely due to the lack of techniques capable of providing detailed information about the arrangement of magnetic moments near interfaces. Here we present polarization-dependent X-ray magnetic dichroism spectro-microscopy that reveals the micromagnetic structure on both sides of a ferromagnetic–antiferromagnetic interface. Images of thin ferromagnetic Co films grown on antiferromagnetic LaFeO₃ show a direct link between the arrangement of spins in each material. Remanent hysteresis loops, recorded for individual ferromagnetic domains, show a local exchange bias. Our results imply that the alignment of the ferromagnetic spins is determined, domain by domain, by the spin directions in the underlying antiferromagnetic layer.

We investigated a thin Co film on top of a 40 nm LaFeO₃ film grown on SrTiO₃(001). The sample was prepared in a molecular beam epitaxy system with the LaFeO₃ film grown using a block-by-block growth method⁶ at 750 °C under a beam of atomic oxygen and a partial O₂ pressure of 5×10^{-6} Torr. This method has been shown to yield high-quality epitaxial films⁷. Plan-view electron-diffraction

and conventional transmission electron microscopy (TEM) analysis show that the epitaxial LaFeO₃ film consists of two microscopic crystallographic domains characterized by orientations of the LaFeO₃ c-axis along the [100] and [010] directions in the SrTiO₃ surface plane⁸. The Co film, grown in the form of a stepped wedge, was deposited *in situ* at room temperature and capped with a 1-nm Pt layer to prevent its oxidation. X-ray diffraction analysis of a sample with a Co thickness of 2.5 nm (and a 1-nm Pt cap layer) revealed a polycrystalline Co structure. Kerr measurements showed that the easy magnetization axis of the sample was in-plane with uniaxial symmetry about the surface normal. All measurements reported here were performed on as-grown samples. Because the samples were not set in a magnetic field they did not exhibit a macroscopic exchange bias^{4,5}. (For LaFeO₃, chemical decomposition prevents the use of setting temperatures close to the Néel temperature (740 K) to create large exchange bias. However, low-temperature (390 K) annealing of Co/LaFeO₃ in a 500-Oe applied field is sufficient to produce a macroscopic bias of 0.007 erg cm⁻² at room temperature and 0.1 erg cm⁻² at 4.2 K). Our ability to probe the spatially resolved magnetic configuration allowed us to observe local bias effects, which average to zero macroscopically.

Spectro-microscopy studies were carried out using the PEEM2 facility at the Advanced Light Source in Berkeley⁹. The focused X-rays, whose polarization could be changed from linear to right or left circular¹⁰, are incident on the sample at an angle of 30° from the surface and form a 30-μm spot. The low-energy secondary photoelectrons from the sample are imaged by an all-electrostatic photo-emission electron microscope (PEEM) with magnification onto a phosphor screen that is read by charge-coupled device (CCD) camera. The spatial resolution of PEEM2 is limited by chromatic aberrations to 20 nm. For imaging we exploited several unique spectroscopic capabilities associated with the variable energy and polarization of the X-rays¹⁰. By tuning the photon energy to either the Fe L-edge, near 710 eV, or the Co L-edge, near 780 eV, we can record separate images of the antiferromagnetic (AFM) LaFeO₃ layer and the ferromagnetic (FM) Co layer. We use linear X-ray

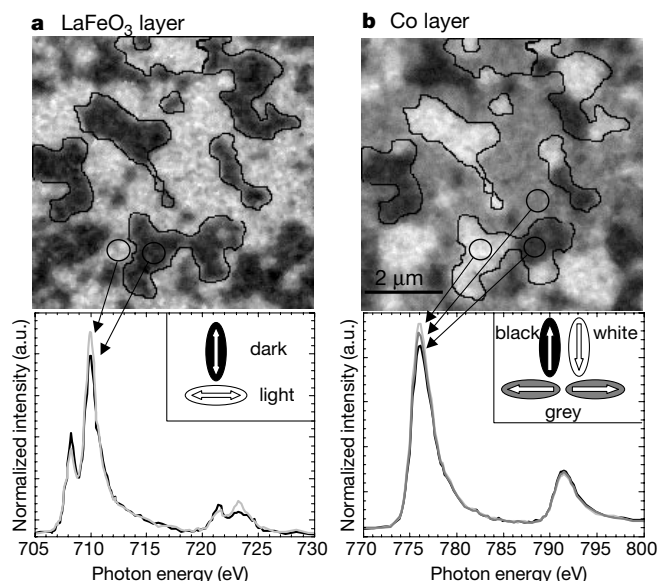


Figure 1 Images and local spectra from the antiferromagnetic and ferromagnetic layers for 1.2-nm Co on LaFeO₃/SrTiO₃(001). **a**, Fe L-edge XMD image; **b**, Co L-edge XMCD image. The contrast in the images arises from antiferromagnetic domains in LaFeO₃ (**a**) and ferromagnetic domains in Co (**b**) with in-plane orientations of the antiferromagnetic axis and ferromagnetic spins as indicated below the images. The spectra shown underneath were recorded in the indicated areas and illustrate the origin of the intensity contrast in the PEEM images.

polarization to image the microscopic AFM domain structure for LaFeO_3 , making use of the large X-ray magnetic linear dichroism (XMLD) effect associated with the multiplet structure at the Fe L_3 or L_2 edge⁸. In our PEEM geometry the electric field vector $\mathbf{E}$ is oriented parallel to the surface. We use right- or left-handed circular polarization to image the FM Co domain structure, exploiting the X-ray magnetic circular dichroism (XMCD) effect at the Co L_3 and L_2 edges¹¹. In our case, the photon angular momentum was oriented parallel to the X-ray propagation direction, at a 30° angle from the surface. Finally, the relatively short sampling depth (about 2 nm) of the PEEM technique¹⁰ combined with the small thickness of the Co layer allows us to be sensitive to the FM and AFM structure in the vicinity of the interface.

Figure 1 shows images of the domain structure in the AFM LaFeO_3 film and in a 1.2-nm-thick FM Co layer on top of the very same substrate region. The magnetic contrast in the left image arises from AFM domains in LaFeO_3 with an in-plane projection of the AFM axis $\mathbf{A}$ oriented horizontally (light) and vertically (dark). The four AFM domains in epitaxial LaFeO_3 have their AFM axes tilted by 45° from the surface normal with orthogonal in-plane projections⁸. As in our experimental geometry $\mathbf{E}$ lies in the surface plane, we cannot distinguish the two domains with collinear in-plane projections and we only observe two of the four AFM domains. Comparison of TEM and PEEM images shows that the AFM domains are seeded by the two epitaxial crystallographic domains⁸. XMLD spectra recorded in the light and dark regions, shown underneath, reveal the spectroscopic origin of the AFM contrast. The FM Co image (Fig. 1b) exhibits three distinct grey scales, corresponding to FM domains aligned vertically up (black) and down (white), and horizontally left or right (grey). For the experimental geometry used for Fig. 1, corresponding to a vertical X-ray propagation direction (and angular momentum), we cannot distinguish left from right horizontally oriented FM domains¹⁰, but these were resolved by a 90° rotation of the sample (not shown). XMCD spectra recorded for regions with different grey scales (Fig. 1b) illustrate the origin of the intensity contrast. Comparison of the in-plane projections of the

AFM axis and the FM spin directions, illustrated below the images, reveals that the FM Co spins are aligned parallel or antiparallel to the in-plane projection of the AFM axis. The correlation revealed by Fig. 1 is of magnetic rather than crystallographic origin as the Co is polycrystalline. The magnetic alignment of the Co domains, which exhibit an in-plane easy axis, must therefore be caused by a coupling to uncompensated spins at the LaFeO_3 surface with an in-plane component parallel to the in-plane projection of the AFM axis. Experiments that attempted to image the uncompensated Fe spins directly, with XMCD microscopy, were unsuccessful, probably because of their small concentration or magnitude. The uncompensated spins may originate solely from an interface effect as discussed in ref. 12 or from the small parasitic Fe magnetization in LaFeO_3 (ref. 13).

The exchange coupling at the Co/LaFeO_3 interface also causes a local exchange bias in individual Co domains. In Fig. 2a a series of XMCD images demonstrate the magnetization reversal process in the Co layer. The measurements were performed on a 2.5-nm Co/LaFeO_3 sample with the magnetic field applied along the X-ray propagation direction. The microscopic XMCD images were acquired in zero field, because of the strong deflection of the photo-electrons in an applied in-plane field. The images reveal a strong uniaxial anisotropy of the Co domains. Only those Co domains that are coupled to LaFeO_3 domains with the projection of the AFM axis parallel to the field exhibit a magnetization reversal (black to white) while the orthogonal Co domains (grey) follow a hard axis loop and remain unchanged. We furthermore observe a local unidirectional magnetic anisotropy of single Co domains, a local exchange bias. Local remanent hysteresis loops (Fig. 2b), calculated from the field dependent XMCD contrast in a series of images, show a repeatable loop shift up to 30 Oe. This local bias is attributed to a surplus of uncompensated spins in the individual AFM domains which are frozen in after growth. (The dipole fields from the neighbouring Co domains are too small to cause this local bias. Micromagnetic calculations showed that these dipole fields are below 5 Oe for a typical domain pattern (domain size between $0.8 \mu\text{m}^2$ and $1.5 \mu\text{m}^2$)). Averaged over the total area shown in Fig. 2a, the bias effects of individual domains cancel out, leading to the unshifted averaged hysteresis loop, also shown in Fig. 2b. The averaged loop exhibits only half of the XMCD brightness change because the orthogonal grey domains constitute half of the signal but do not switch. The lack of bias in the averaged loop is expected, because the studied sample was not set in a magnetic field. The setting procedure would shift the balance of the microscopically biased domains, resulting in a preferred macroscopic spin direction, that is, exchange bias.

Our results open the door for investigations of various ferromagnetic–antiferromagnetic systems consisting of single crystal, polycrystalline or even amorphous materials. As such, they may hold the key to a definitive understanding of the ferromagnetic–antiferromagnetic exchange bias phenomenon. □

Received 10 November 1999; accepted 5 April 2000.

- Meiklejohn, W. H. & Bean, C. P. New magnetic anisotropy. *Phys. Rev.* **102**, 1413–1414 (1956).
- Tsang, C. *et al.* Design, fabrication and testing of spin-valve read heads for high density recording. *IEEE Trans. Mag.* **30**, 3801–3806 (1994).
- Parkin, S. S. P. *et al.* Exchange-biased magnetic tunnel junctions and application to non-volatile magnetic random access memory. *J. Appl. Phys.* **85**, 5828–5833 (1999).
- Nogués, J. & Schuller, I. K. Exchange bias. *J. Magn. Magn. Mater.* **192**, 203–232 (1999).
- Berkowitz, A. E. & Takano, K. Exchange anisotropy—a review. *J. Magn. Magn. Mater.* **200**, 552–570 (1999).
- Locquet, J.-P., Catana, A., Mächler, E., Gerber, C. & Bednorz, J. G. Block-by-block deposition: A new growth method for complex oxide thin films. *Appl. Phys. Lett.* **64**, 372–374 (1994).
- Locquet, J.-P. *et al.* Doubling the critical temperature of $\text{La}_{1-x}\text{Sr}_x\text{CuO}_4$ using epitaxial strain. *Nature* **394**, 453–456 (1998).
- Scholl, A. *et al.* Observation of antiferromagnetic domains in epitaxial thin films. *Science* **287**, 1014–1016 (2000).
- Anders, S. *et al.* Photoemission electron microscope for the study of magnetic materials. *Rev. Sci. Instrum.* **70**, 3973–3981 (1999).
- Stöhr, J., Padmore, H. A., Anders, S., Stämmler, T. & Scheinfein, M. R. Principles of x-ray magnetic dichroism spectromicroscopy. *Surf. Rev. Lett.* **5**, 1297–1308 (1998).

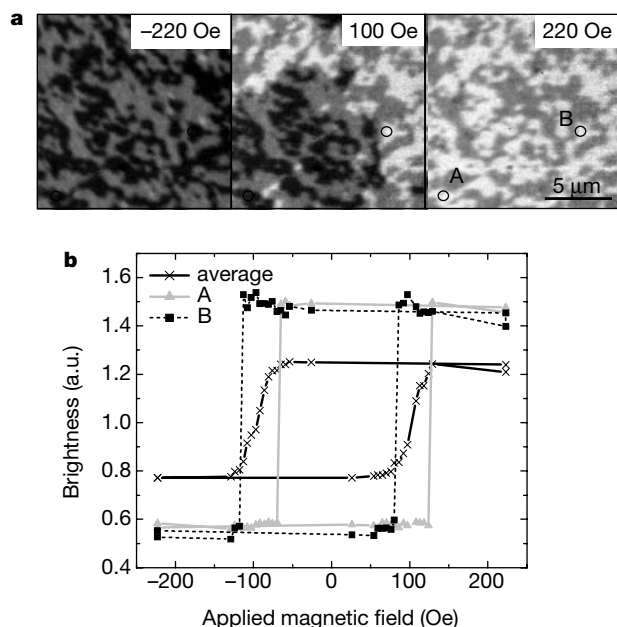


Figure 2 Field dependence of Co domains for a 2.5-nm Co/LaFeO_3 sample. **a**, Images of the Co domain pattern after applying a magnetic field of -220 Oe, 100 Oe and 220 Oe. The field was applied along the X-ray polarization direction and the images were acquired in zero field. **b**, Local remanent hysteresis loops calculated from the XMCD contrast for two different areas A and B inside domains with the easy axis along the field direction (black and white) and for the entire area in **a**, including grey domains which have their easy axis perpendicular to the field direction.

11. Stöhr, J. *et al.* Element specific magnetic microscopy with circularly polarized X-rays. *Science* **259**, 658–661 (1993).
12. Takano, K. *et al.* Interfacial uncompensated antiferromagnetic spins: role of unidirectional anisotropy in polycrystalline $\text{Ni}_{81}\text{Fe}_{19}/\text{CoO}$ bilayers. *Phys. Rev. Lett.* **79**, 1130–1133 (1997).
13. Eibschütz, M., Shtrikman, S. & Treves, D. Mössbauer studies of Fe^{57} in orthoferrites. *Phys. Rev.* **156**, 562–577 (1967).

Acknowledgements

This work was supported by the Director, Office of Basic Energy Sciences, of the US Department of Energy. J.W.S. and E.N. acknowledge support by the Swiss National Science Foundation.

Correspondence and requests for materials should be addressed to E.N. (e-mail: FNolting@lbl.gov).

Reversible electromechanical characteristics of carbon nanotubes under local-probe manipulation

Thomas W. Tombler*, Chongwu Zhou*, Leo Alexseyev*, Jing Kong*, Hongjie Dai*, Lei Liu†, C. S. Jayanthi‡, Meijie Tang‡ & Shi-Yu Wu†

* Department of Chemistry, Stanford University, Stanford, California 94305, USA

† Department of Physics, University of Louisville, Louisville, Kentucky 40292, USA

‡ Physics Directorate, Lawrence Livermore National Laboratory, Livermore, California 94551, USA

The effects of mechanical deformation on the electrical properties of carbon nanotubes are of interest given the practical potential of nanotubes in electromechanical devices, and they have been studied using both theoretical^{1–4} and experimental^{5,6} approaches. One recent experiment⁶ used the tip of an atomic force microscope (AFM) to manipulate multi-walled nanotubes, revealing that changes in the sample resistance were small unless the nanotubes fractured or the metal–tube contacts were perturbed. But it remains unclear how mechanical deformation affects the intrinsic electrical properties of nanotubes. Here we report an experimental and theoretical elucidation of the electromechanical characteristics of individual single-walled carbon nanotubes (SWNTs) under local-probe manipulation. We use AFM tips to deflect suspended SWNTs reversibly, without changing the contact resistance; *in situ* electrical measurements reveal that the conductance of an SWNT sample can be reduced by two orders of magnitude when deformed by an AFM tip. Our tight-binding simulations indicate that this effect is owing to the formation of local sp^3 bonds caused by the mechanical pushing action of the tip.

We prepared samples of individual SWNTs bridging metal electrodes on SiO_2/Si substrates^{7–9}, with part of the SWNT length suspended over trenches fabricated on the SiO_2 surface (Fig. 1a). We characterized the partially suspended individual SWNTs by AFM imaging. Figure 1b shows the AFM image of an SWNT with suspended length $l \approx 605$ nm over a trench. The image was obtained by tapping-mode AFM (TM-AFM) with the tip scanning direction parallel to the tube axis. Figure 1c shows the experimental setup for bending a suspended SWNT mechanically with an AFM tip while measuring the nanotube electrical properties.

After a desired SWNT device was located and imaged by TM-AFM, the AFM tip was positioned above the centre of the suspended nanotube. The nanotube suspension was pushed towards the bottom of the trench by moving the sample-stage upward. The stage was then retracted. The pushing–retracting cycle was repeated

many times, during which the AFM cantilever deflection and the resistance of the SWNT sample were simultaneously recorded as a function of time. This approach allowed repeated measurements of resistance versus nanotube deflection, as the cantilever deflection signal (ΔZ_c) can readily be converted into the deflection of the suspended nanotube at its centre point (ΔZ_T , Fig. 1c). By controlling the initial tip–tube distance Z_0 and the total sample-stage travel range (Z_{range}), we were able to deflect the suspended SWNT to various degrees and study the effect of mechanical deformation on the electrical properties of SWNTs.

Figure 2 shows cantilever deflection ΔZ_c versus vertical coordinate Z recorded during one cycle of pushing on the suspended SWNT sample shown in Fig. 1b. Beyond the tip–tube contact point $Z_0 \approx 50$ nm, the vertical deflection occurring at the centre of the suspended SWNT is $\delta(Z) = \Delta Z_T(Z) = (Z - Z_0) - \Delta Z_c(Z)$, and the force applied to the nanotube is $F(\delta) = k_c \Delta Z_c(\delta)$. We find that the force $F(\delta)$ versus nanotube deflection δ curve (Fig. 2, inset) right after the tube–tip contact can be fitted well into $F(\delta) = 8YA(\delta/l)^3$,

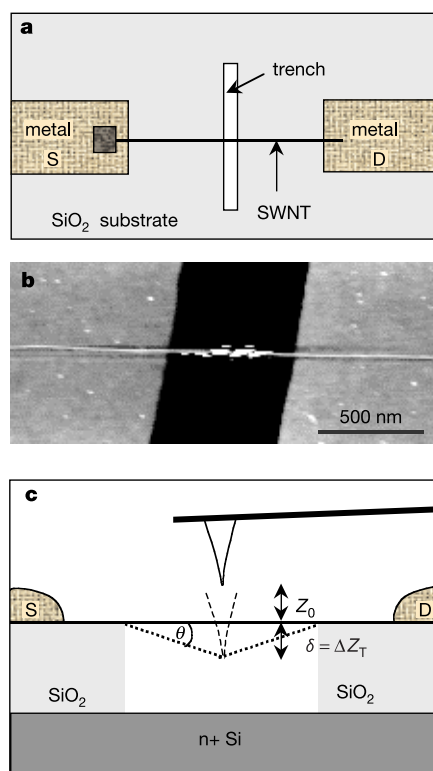


Figure 1 An SWNT partly suspended over a trench for electromechanical measurements. **a**, Device viewed from above. Preparation of samples involves chemical vapour deposition of SWNTs at desired surface sites using SiO_2/Si substrates with patterned catalyst islands^{7–9}. The substrates contain trenches that are about 500 nm wide and 175 nm deep, pre-fabricated next to patterned catalyst islands (dark square). Thus, the SWNT bridging a pair of metal electrodes (S is the source, D is the drain) is partly suspended over the trench. The spacing between metal electrodes is about 3–4 μm . The metal used to contact SWNTs is 20 nm thick Ti and 60 nm Au placed on top of the SWNTs over a contact length of about 1 μm . **b**, AFM image of an SWNT with suspended length $l \approx 605$ nm. The cantilever employed for this experiment has a spring constant $k_c = 0.6 \text{ N m}^{-1}$. The integrated tip on the cantilever is pyramidal with tip radius of about 10–15 nm. The bright streaks around the suspended tube are caused by tube touching and sticking to the side of the pyramid when the tip is scanned near the tube. The diameter of the SWNT $d = 3.1 \pm 0.2$ nm, measured from the apparent height of the nanotube resting on the SiO_2 surface. The nanotube is a relatively large diameter SWNT synthesized by our chemical vapour deposition approach⁷. It could also be a small SWNT bundle, but this should not change our main conclusions. **c**, Side-view of the AFM pushing experiment. The tip is centred above the SWNT suspension by slowly zooming into the tube-suspension during real-space imaging.

where $A = (\pi d)t$ (where $t = 3.4 \text{ \AA}$, van der Waals wall thickness) is the nanotube cross-section area. This yields a Young's modulus of the nanotube $Y \approx 1.2 \text{ TPa}$, similar to previous results $Y = 0.6 - 1.3 \text{ TPa}$ for SWNTs¹⁰. The $F(\delta) \propto \delta^3$ fitting indicates that the suspended SWNT can be considered as an elastic string when it is under initial force loading at the centre^{11,12}. Assuming that the deflected SWNT is pivoted at the edges of the trench and forms a triangle with its original configuration (Fig. 1c), we define $\sigma = [(4\delta^2 + l^2)^{1/2} - l]/l$ as a global strain parameter distributed over the length of the nanotube.

Alternatively, a parameter $\theta = \tan^{-1}(2\delta/l)$ is defined as the angle between the deflected SWNT and its original configuration (Fig. 1c). For $\sigma > \sim 0.3\%$ ($\theta > \sim 5^\circ$), the force versus deflection relation is found to deviate from $F(\delta) \propto \delta^3$ (Fig. 2).

Figure 3 shows the cantilever deflection and sample conductance evolution during repeated tip-pushing of the suspended part of the SWNT. For initial tip-tube distances $Z_0 \approx 65 \text{ nm}$, 30 nm and 8 nm and sample-stage moving range $Z_{\text{range}} = 100 \text{ nm}$, the maximum cantilever and tube deflections are $\Delta Z_c \sim 3, 10$ and 15 nm , and

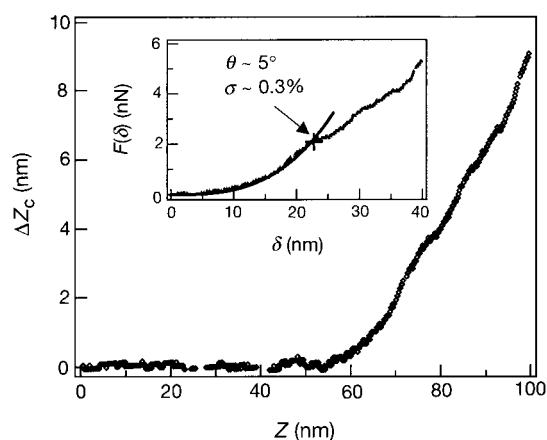


Figure 2 Cantilever deflection ΔZ_c versus vertical coordinate Z during a cycle of pushing a suspended nanotube and then retracting. Inset, force $F(\delta)$ versus nanotube deflection δ curve. Solid line: fit of $F(\delta) = 8YA(\delta/l)^3 \approx 1.53 \times 10^{-14} (\text{N m}^{-3})\delta^3$. The arrow highlights the deviation point from $F \propto \delta^3$ where the tube bending angle $\theta \approx 5^\circ$ and global strain $\sigma \approx 0.3\%$. We carried out the force-deflection analysis by first estimating a tip-tube

contact point Z_0 from the ΔZ_c versus Z curve. A final Z_0 value was determined by varying Z_0 by small increments until reaching the best $F \propto \delta^3$ fit for the initial tube-tip contact regime. Such analysis allowed the determination of strain σ and bending angle θ from the cantilever deflection signal. This is useful for the characterization of nanotube electro-mechanical properties (see Figs 3 and 4).

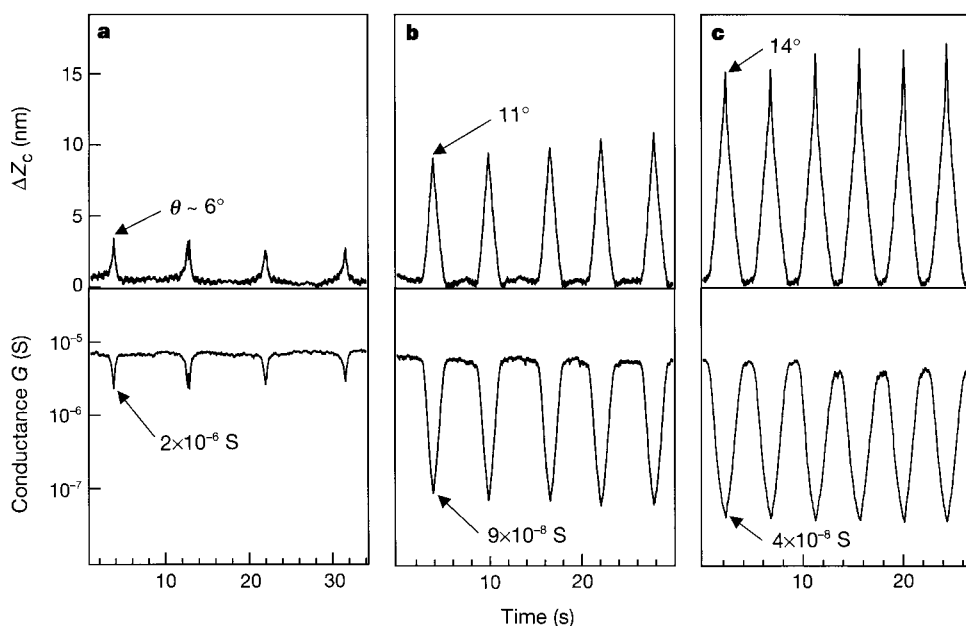
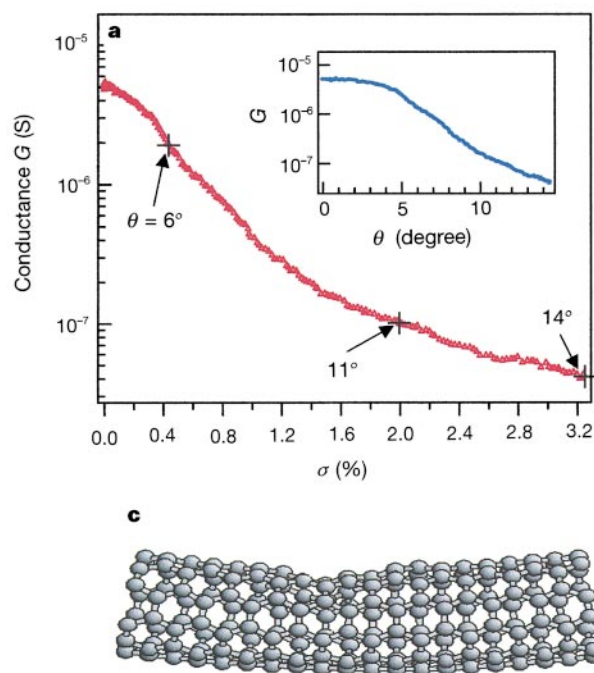


Figure 3 Cantilever deflection and nanotube electrical conductance evolution during repeated cycles of pushing the suspended SWNT. Initial tip-tube distance is $Z_0 \approx 65 \text{ nm}$ (a), 30 nm (b) and 8 nm (c). The speed of the tip motion was set to be $\sim 22 \text{ nm s}^{-1}$ (a), 34 nm s^{-1} (b) and 44 nm s^{-1} (c). Tip travel distance = $2Z_{\text{range}} = 200 \text{ nm}$ between oscillation peaks. Time resolution for the conductance measurements was 10 ms . The arrows point to the bending angles and conductance values for the first peak. The bending angles were obtained by force-deflection analysis of the three marked pushing-retracting cycles independently, using the method described in Fig. 2. The angles for the marked peaks in a and b were also evaluated based on the analysis of the marked peak in c. This was done by correlating the cantilever deflection values for the two peaks in a and

b with the cantilever deflection versus angle curve obtained from the analysis of the marked cycle in c. The results were found to be consistent with the independent analyses of the marked cycles in a and b. Drifts in Z_0 (due to piezo-drifts) occurred between some of the pushing-retracting cycles, but within each cycle the drift should not be significant. Pushing the SWNT to its breaking point would provide information such as nanotube tensile strength and electrical behaviour near tube-failure. This is currently limited by the size of the trench (500 nm wide, 175 nm deep) on the samples. Note that the nanotube when unperturbed exhibited a resistance of about 200 kilo-ohms . This could be mainly contributed by contact resistance.

$\delta = \Delta Z_T \sim 30, 60$ and 76 nm, respectively (Fig. 3a–c). The conductance of the SWNT sample decreases each time the AFM tip pushes the nanotube down, but recovers as the tip retracts. The repeated pushing–retracting causes oscillations in the cantilever–nanotube deflection and sample conductance, with equal periodicity in the two oscillations. Importantly, both the mechanical deformation and electrical conductance of the nanotube are highly reversible. The full reversal of these characteristics upon tip retraction has important implications. First, reversibility in the electrical property indicates that the metal–tube contacts are not affected when the tip deflects the suspended part of the nanotube. The observed change in sample conductance is entirely owing to the mechanical deformation of the SWNT caused by the pushing tip. Secondly, reversibility in both the mechanical and electrical properties indicates that the suspended part of the SWNT is firmly pivoted at the edges of the trench. The length of the SWNT resting on the SiO_2 surface ($\geq 1.5 \mu\text{m}$ on both sides of the trench) is not being stretched or sliding on the substrate during pushing of the suspended part. Anchoring of the nanotube on the SiO_2 substrate is attributed to strong tube–substrate van der Waals interactions¹³.

The conductance of the SWNT sample decreases by more than two orders of magnitude when the AFM tip deflects the SWNT by $\delta = \Delta Z_T \approx 80$ nm (Fig. 3c). This deflection corresponds to a global strain value $\sigma \approx 3\%$ ($\theta = 14^\circ$) in the nanotube suspension. Conductance of the SWNT sample as continuous functions of strain and bending angle are shown in Fig. 4a. The SWNT conductance decreases relatively slowly for small bending angles ($\theta \leq 5^\circ$), and the decrease becomes more dramatic at higher bending angles (Fig. 4a, inset).



To understand the physics underlying the dramatic nanotube electromechanical property, we carried out order-N non-orthogonal tight-binding molecular-dynamics simulations¹⁴ of an AFM tip deflecting a (5,5) SWNT (960 carbon atoms, $l \approx 12$ nm). Electrical measurements with the SWNT sample described in Fig. 3 identified that the nanotube was metallic in nature, as the sample exhibited little conductance change under various back-gate voltages and remained conducting at low temperatures^{5,8,9}. Therefore, we chose the (5,5) metallic tube for our simulation. The AFM tip was modelled by a capped (5,5) SWNT (110 carbon atoms). The tip was placed directly above the middle of the suspended SWNT and pushed down vertically. The simulations were carried out at 300 K. In the simulation, 40 atoms at each end of the SWNT and the 50 atoms at the top end of the tip were kept fixed. The rest of the atoms in the system were allowed to relax and reach equilibrium configuration. The conductance of the bent SWNT was calculated by the Landauer–Büttiker formula¹⁵ using the method of real space Green's function¹⁶. Details of the simulations and conductance calculations will be presented elsewhere. We calculated the conductance evolution of the nanotube as it was deflected to $\theta = 0^\circ, 7.0^\circ, 11^\circ$, and 15° , corresponding to $\sigma = 0, 0.7\%, 1.8\%$, and 3.4% , respectively (Fig. 4b). The conductance (in units of $G_0 = 2e^2/h$) at the Fermi level was found to decrease by about twofold (from $2 G_0$ to $\sim 1 G_0$) at $\theta = 7^\circ$, and decrease more dramatically at larger bending angles ($G \approx 0.01 G_0$ at $\theta = 15^\circ$). These results are qualitatively consistent with our experimental data (Fig. 4a, inset).

Detailed analysis of our simulation results reveals that strong local bonding deformation induced by the AFM tip is responsible for the dramatic conductance decrease of the SWNT. When pushed

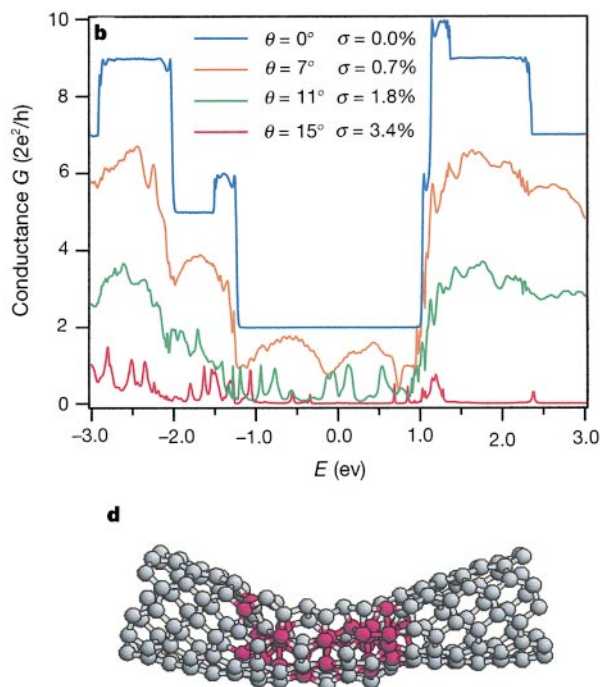


Figure 4 Electrical conductance versus mechanical deformation for a manipulated SWNT. **a**, Experimental result of conductance (G) of the SWNT sample versus strain (σ) in the suspended part of the nanotube. Inset, conductance (G) versus bending angle (θ). These data were obtained from the marked cycle in Fig. 3c. The strain (σ) does not scale linearly with θ , which accounts for the different shapes of the G – σ and G – θ curves.

b, Conductance versus band energy calculated for an ideally contacted (5,5) SWNT under progressively larger deformation forced by a tip. The Fermi energy is at $E = 0$. The fluctuations of conductance versus energy (around E_F , for $\theta \neq 0$) are owing to variations of the transmission coefficient for different energy electrons crossing the barrier set by the large local deformation region near the tip. **c,d**, Simulated atomic configurations of the nanotube pushed to 7° and 15° respectively. Atoms marked in red are near the tip (not

shown) and sp^3 -bonded. The sp^3 bonding is found to be within the tube structure itself (between back-and-belly atoms in the nanotube). Bond number counting excludes tip-atoms. The role of the tip is to act as a forcing agent. The tip used in the experiment can be considered as a sharp point relative to the length of the tube. Also, the AFM tips typically do not have smooth textures, which may have led to a situation where the tube–tip contact is at a sharp irregular point on the tip. We measured several nanotube samples with different tips, and observed similar electromechanical properties of suspended metallic SWNTs. For instance, another SWNT (diameter ~ 1.9 nm, suspended length ~ 350 nm) manipulated by a different AFM tip exhibited a conductance decrease from 7×10^{-6} S to 3×10^{-6} S under $\sigma \approx 0.3\%$ and $\theta \approx 4^\circ$.

by an AFM tip, the nanotube region proximal to the tip exhibits significant changes in atomic bonding configuration. For relatively small bending angles ($\theta \leq \sim 7^\circ$), the nanotube retains sp^2 bonding throughout its structure. The nanotube responds elastically but exhibits a larger bond distortion (up to 5.5%) for the atoms in the region underneath the tip than the global strain ($\leq 0.7\%$). This distortion accounts for the initial conductance decrease at small bending angles where the overall simulated structure remains sp^2 . As tip-pushing proceeds, the tube structure progressively evolves and larger structural changes occur underneath the tip. The average number of bonds per atom in the tube section proximal to the tip has increased from 3 to ~ 3.3 for deflection angle $\theta = 11^\circ$ ($\sigma = 1.8\%$) and to 3.6 for $\theta = 15^\circ$ ($\sigma = 3.4\%$). This indicates that the local bonding configuration has changed from sp^2 to nearly sp^3 . Away from the tip region, the nanotube remains essentially in a sp^2 -bonding configuration with bond deformation characterized by the global strain parameter σ . Further local analysis¹⁷ of the bent nanotube in the tip vicinity reveals the onset of an increase in σ -electrons contributing to the local density of states at the Fermi level in the highly deformed local region. This is, however, accompanied by a significant decrease in the π -electron density. As the π -electrons are delocalized and therefore mainly responsible for electrical conduction, a drastic reduction in the π -electron density is responsible for the substantial decrease in conductance. Simulations also find that the local and global deformations of the nanotube are highly reversible (for $\theta < 15^\circ$, $\sigma < 3.4\%$) upon moving the tip away, leading to the recovery of the nanotube structure and electrical conductance. These results are in excellent agreement with our experimental observations.

The deviation of force versus deflection $F(\delta)$ curve from the δ^3 -relation for $\sigma > \sim 0.3\%$ ($\theta > \sim 5^\circ$, Fig. 2) should be owing to large local-strain developed in the sp^3 region proximal to the tip, as the elastic string model assuming a homogeneous global strain becomes invalid. This is consistent with the electromechanical behaviour that beyond the elastic response regime ($\theta > \sim 5^\circ$, Fig. 4a, inset), tip-forced sp^3 bonding within the nanotube occurs, causing a significant decrease in the nanotube conductance.

Previous theoretical investigations indicate that the electrical properties of metallic SWNTs should be insensitive to small bending deformations^{2,4,18}. The calculated conductance of a (5,5) SWNT changes very little for bending angles up to $\theta = 24^\circ$, when bending at the centre of the SWNT is modelled by holding the ends of the tube at fixed positions to define the bending angle without the involvement of a tip³. This results in a situation where the atomic bonding characteristics of the bent nanotube still remain sp^2 . The absence of sp^3 bonding in the simulated structure should account for the small electrical conductance change³. A similar bending technique¹⁸ found that the electrical conductance of a metallic (6,6) SWNT did not change significantly for bending angles up to $\theta \sim 22.5^\circ$. At larger bending angles (for example $\theta = 45^\circ$) the conductance of the SWNT was lowered by at most tenfold. The conductance decrease is explained by σ - π hybridization effects owing to the increased curvature under high bending angles¹⁸.

Here we used an AFM tip both in experiments and in simulations, a key to performing the experimental measurements and obtaining a fundamental understanding. Our work elucidates the electromechanical properties of the nanotube when mechanical action of a local probe causes a large local deformation. This differs from previous considerations of deformed nanotubes in which the nanotube structure is more or less uniformly bent or strained (at least in the bending-angle range investigated here). The experimental investigation of uniform global strain or bending effects requires different nanotube manipulation mechanisms, such as electrostatic forces¹⁹. We believe that the physics presented here should hold for SWNTs containing large local deformations caused by other forces. For instance, if a highly kinked SWNT stabilized by van der Waals forces on a substrate develops sp^3 bonding character-

istics at the kink, the electrical conductance should be significantly reduced compared to a straight tube. □

Received 8 December 1999; accepted 20 April 2000.

- Crespi, V., Cohen, M. & Rubio, A. *In situ* band gap engineering of carbon nanotubes. *Phys. Rev. Lett.* **79**, 2093–2096 (1997).
- Kane, C. L. & Mele, E. J. Size, shape, and low energy electronic structure of carbon nanotubes. *Phys. Rev. Lett.* **78**, 1932–1935 (1997).
- Nardelli, M. & Bernholc, J. Mechanical deformations and coherent transport in carbon nanotubes. *Phys. Rev. B*, **60**, R16338–R16341 (1998).
- Rocheffort, A., Salahub, D. & Avouris, P. The effect of structural distortions on the electronic structure of carbon nanotubes. *Chem. Phys. Lett.* **297**, 45–50 (1998).
- Bezryadin, A., Verschuere, A., Tans, S. & Dekker, C. Multiprobe transport experiments on individual single-wall carbon nanotubes. *Phys. Rev. Lett.* **80**, 4036–4039 (1998).
- Paulson, S. *et al.* *In situ* resistance measurements of strained carbon nanotubes. *Appl. Phys. Lett.* **75**, 2936–2938 (1999).
- Kong, J., Soh, H., Cassell, A., Quate, C. F. & Dai, H. Synthesis of individual single-walled carbon nanotubes on patterned silicon wafers. *Nature* **395**, 878–881 (1998).
- Soh, H. *et al.* Integrated nanotube circuits: controlled growth and ohmic contacting of single-walled carbon nanotubes. *Appl. Phys. Lett.* **75**, 627–629 (1999).
- Kong, J. *et al.* Synthesis, integration and electrical properties of individual single-walled carbon nanotubes. *Appl. Phys. A* **69**, 305–308 (1999).
- Salvetat, J. P. *et al.* Elastic and shear moduli of single-walled carbon nanotube ropes. *Phys. Rev. Lett.* **82**, 944–947 (1999).
- Timoshenko, S. *Strength of Materials. Part I, Elementary Theory and Problems* (Van Nostrand, New York, 1930).
- Walters, D. *et al.* Elastic strain of freely suspended single-walled carbon nanotube ropes. *Appl. Phys. Lett.* **74**, 3803–3805 (1999).
- Hertel, T., Martel, R. & Avouris, P. Manipulation of individual carbon nanotubes and their interaction with surfaces. *J. Phys. Chem.* **102**, 910–915 (1998).
- Jayanthi, C. S. *et al.* Order-N method for a nonorthogonal tight-binding Hamiltonian. *Phys. Rev. B* **57**, 3799–3802 (1998).
- Datta, S. *Electronic Transport in Mesoscopic Systems*. (University Press, Cambridge, 1995).
- Wu, S. Y. & Jayanthi, C. S. Local analysis via the real space greens function method. *J. Modern Phys.* **B 9**, 1869–1897 (1995).
- Alfonso, D., Wu, S. Y., Jayanthi, C. S. & Kaxiras, E. Linking chemical reactivity, magic numbers, and local electronic properties of clusters. *Phys. Rev. B* **59**, 7745–7750 (1999).
- Rocheffort, A., Lesage, F., Salhub, D. & Avouris, P. Conductance of distorted carbon nanotubes. *Phys. Rev. B* **60**, 13824–13830 (1999).
- Kim, P. & Lieber, C. Nanotube nanotweezers. *Science* **286**, 2148–2150 (1999).

Acknowledgements

We thank C. Quate for discussions and use of equipment. This work was supported by the Defense Advanced Research Projects Agency/Office of Naval Research, National Science Foundation, Semiconductor Research Corporation/Motorola, a David and Lucile Packard Fellowship, a Terman Fellowship, the Laboratory for Advance Materials at Stanford, National Nanofabrication Users Network at Stanford, the Camille Henry-Dreyfus Foundation, the American Chemical Society and the University of Kentucky Center for Computer Sciences.

Correspondence and requests for materials should be addressed to H.D. (email: hda@chem.stanford.edu).

Controlling droplet deposition with polymer additives

Vance Bergeron[†], Daniel Bonn[‡], Jean Yves Martin[†] & Louis Vovelle[†]

[†] Rhodia Recherches, Centre de Recherches Lyon, 85 Av. Des Freres Perret-BP62, 69192 Saint-Fons Cedex, France

[‡] Ecole Normale Supérieure, Laboratoire de Physique Statistique, 24 rue Lhomond, 75231 Paris Cedex 05, France

Controlling the impact of drops onto solid surfaces is important for a wide variety of coating and deposition processes—for example, the treatment of plants with herbicides and pesticides requires precise targeting in order to meet stringent toxicological regulations. However, the outer wax-like layer of the leaves is a non-wetting substrate that causes sprayed droplets to rebound; often less than 50% of the initial spray is retained by the plant¹. Although the impact and subsequent retraction of non-wetting aqueous drops on a hydrophobic surface have been the subjects of

extensive experimental and theoretical work^{2–7}, non-newtonian rheological effects have not been considered in any detail. Here we report that, by adding very small amounts of a flexible polymer to the aqueous phase, we can inhibit droplet rebound on a hydrophobic surface and markedly improve deposition without significantly altering the shear viscosity of the solutions. Our results can be understood by taking into account the non-newtonian elongational viscosity, which provides a large resistance to drop retraction after impact, thereby suppressing droplet rebound.

High-velocity drops colliding with a solid non-wetting surface impact, expand and subsequently retract (Fig. 1). Previous work has focused on the very rapid impact and expansion stages in an effort to determine the maximum diameter, $D_{\max}$, that a drop is capable of attaining on impact for optimizing deposition^{2,5,6}. These studies have shown that for newtonian fluids the impact and expansion stages are governed by the Reynolds number, $Re = DV_1\rho/\eta_s$, which is a ratio of inertial to viscous forces, and the Weber number, $We = \rho V_1^2 D/\sigma$, which is a balance between inertial and capillary forces, where D is the drop diameter, V_1 is the drop velocity at impact, ρ is the drop density, η_s is the solution shear viscosity and σ is the surface tension of the solution. In general, the contact angle of the liquid on the substrate and surface roughness must also be considered when describing the impact events of drops.

Here we use only smooth, non-wetting hydrophobic surfaces, all with a receding contact angle greater than 120° . Furthermore, we use large drops (~ 2 mm) striking at high velocities, which leads to inertially dominated impact and expansion stages. This not only corresponds to most practical situations but also allows us to control the first two stages of the impact, that is, that all drops in this study expand to the same maximum diameter⁵. We are therefore able to focus on the final drop-retraction stage of the process which determines whether or not a deposited drop rebounds off the surface.

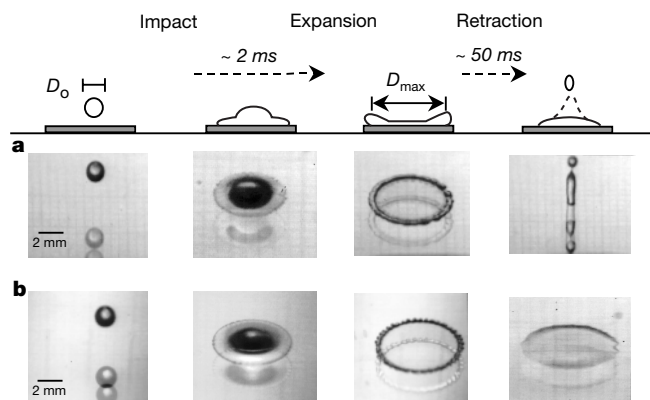


Figure 1 Typical photographic sequence (left to right) of two aqueous drops striking a hydrophobic surface. In each case focus should be placed on the uppermost image; the lower ghost images are reflections from the top and/or bottom surface of the solid glass substrate. Impact and expansion from the initial undeformed drop diameter, D_0 , to the maximum spread diameter, $D_{\max}$, occurs in the first 2 ms. Subsequently, the drops retract and can either detach from the surface (a, pure water) or remain bound to the surface after impact (b, dilute polyethyleneoxide solution, molecular weight 4×10^6 g mol⁻¹, at 0.1 g l⁻¹). We visualize these events with a 1,000 frame s⁻¹ video recording system (NAC HSV-1000). Drops are produced using a syringe pump, extruding the solution through a stainless-steel needle, the diameter of which determines the drop size. Impact velocities are imposed by the height of the needle above the horizontally placed hydrophobic surface. These surfaces are prepared by spin-coating 1×1 cm, acid-washed glass plates with a 0.4 g ml⁻¹ solution of a stearic acid complexing agent, Rhodoline EP 9271 (Rhodia Specialty Chemicals). This produces a smooth coating with an air-solution receding contact angle greater than 120° , both before and after impact, for all of our solutions. This last point verifies that specific adsorption of material (that is, polymer) upon drop impact does not lead to surface modification during the impact event, which is known to influence the wettability of the surface in certain circumstances^{15,16}.

We find that droplet rebound from non-wetting surfaces can be markedly suppressed by small amounts of a polymer additive. High-speed photography reveals similar impact and expansion stages for both pure water and dilute polymer solutions, yet the retraction phase is very different. For pure water (Fig. 1a) the drop retracts violently, leading to ejection of part of the droplet from the surface, that is, drop rebound occurs; however, for dilute aqueous solutions of a flexible polymer (for example, polyethyleneoxide; Fig. 1b) the drop retracts much more slowly and remains deposited on the surface.

The photographic sequences in Fig. 1 are transformed into 'drop-evolution' plots by recording the drop diameter on the surface over time (Fig. 2). These plots permit us to define a retraction velocity, V_{ret} , from which we can quantitatively gauge the drop-retraction phase. Notably, droplet rebound occurs if the retraction speed is beyond a certain critical value, $V_{\text{ret}} > 300$ mm s⁻¹, not attained for the polymeric solutions.

Compared with the inertially dominated impact and expansion stages, which occur over the first 2 ms, the flow during the retraction stage is nearly an order of magnitude slower. This fact is quantitatively revealed by comparing the Reynolds and Weber numbers for pure water droplets during the expansion and retraction stage: $Re_{\text{expansion}}/Re_{\text{retraction}} \approx 6,000/1,200$, $We_{\text{expansion}}/We_{\text{retraction}} \approx 250/10$, the difference resulting from a fivefold decrease in the retraction velocity relative to the drop-impact velocity. It follows that drop inertia dominates during the expansion stage, although, to a good approximation, it can be neglected with respect to viscosity effects during the retraction stage: for newtonian liquids, the retraction speed is governed by the capillary number, $Ca = V_{\text{ret}}\eta_s/\sigma$ (ref. 8), where V_{ret} is the retraction velocity. Ca represents the competition between the viscous forces that tend to slow down the retraction, and the capillary forces (that is, surface tension) that act to contract the drop and accelerate retraction. Thus, systematically comparing the Ca values of a series of aqueous newtonian solutions with those of their non-newtonian counterparts (that is, solutions with the same shear viscosity and surface tension) allows us to test the influence that non-newtonian rheological properties have on droplet rebound.

The dependence of Ca on drop-retraction speed for the case of newtonian fluids is obtained by changing the shear viscosity using water-glycerol mixtures (Fig. 3). Experimental aspects of this case have been previously reported⁹. For dilute polymer solutions the

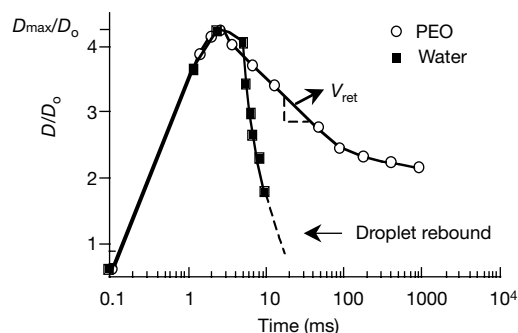


Figure 2 Drop diameter versus time. Filled squares represent a pure water droplet, open circles represent a dilute aqueous solution of polyethyleneoxide (PEO; 0.1 g l⁻¹ of molecular weight 4×10^6 g mol⁻¹); lines are provided as a guide to the eye. Analysis of the full data from all of our high-speed footage shows that to within experimental error the spreading velocities are identical during the impact and expansion stages. Drop-retraction velocity, V_{ret} , is defined as the slope of the spread radius versus time after maximum drop expansion. The characteristic V_{ret} for each drop is determined by making a linear approximation of the initial linear portion of the decaying region of the data. Dashed lines extending from the pure water data indicate that a portion of the droplet has pinched off the surface.

behaviour is different. Here the level of polymer is systematically changed, but the shear viscosity and surface tension for these dilute solutions (between 0.01 and 0.5 g l⁻¹ polyethyleneoxide) remain very similar to those of pure water. The subsequent Ca versus V_{ret} correlation between the newtonian fluids and the dilute polymer solutions diverges. This clearly shows that non-newtonian effects need to be considered. Specifically, owing to the highly elongational nature of the flow within the retracting droplets, the predominant effect is due to the high elongational viscosity, η_e , provided by the flexible polymers in solution (see below).

In determining the resistance to stretching motion in a fluid, the so-called 'elongational' or extensional viscosity has been and remains a significant problem in non-newtonian fluid mechanics in general, and for dilute solutions in particular¹⁰. We obtain the elongational viscosity of our solutions in two different ways: first, by using an opposing-jet rheometer¹¹; and second, by using a technique that exploits the use of hydrodynamic constitutive equations. Fitting an appropriate constitutive equation to direct measurements of the shear viscosity and normal stress difference provides the information needed to evaluate the solution's elongational viscosity. Our measurements are obtained in simple shear flow using a custom-made cone-plate geometry which allows access to the high shear rates that are necessary to measure the elastic response of the dilute polymer solutions. The data are then fitted to the so-called 'FENE-P' (finitely extendible nonlinear elastic) constitutive equation^{12,13}; a relatively simple model in which the polymer molecules are represented as dumb-bells that can be stretched by a finite amount in the flow field. Values of the elongational viscosity obtained in this way are compared with the measurements using a Rheometrics RFX opposing nozzle rheometer referenced to pure water (Fig. 4). Although there has been much discussion concerning the validity of measurements obtained with such opposing-nozzle devices^{10,11}, the agreement with the FENE-P model calculations is satisfactory (see Fig. 4). This is probably because our polymer solutions are very dilute and have shear viscosities close to that of water.

The characteristic elongation rate of the retracting drops can be obtained from the rim velocity (~ 30 cm s⁻¹) divided by the thickness of the stretched droplet (~ 0.1 mm). This leads to typical elongation rates of $\epsilon \approx 3,000$ s⁻¹. With reference to the measurements of the elongational viscosity versus elongation rate (Fig. 4), it is evident that the high- ϵ plateau for η_e should be taken as the characteristic elongational viscosity during our drop-impact experiments.

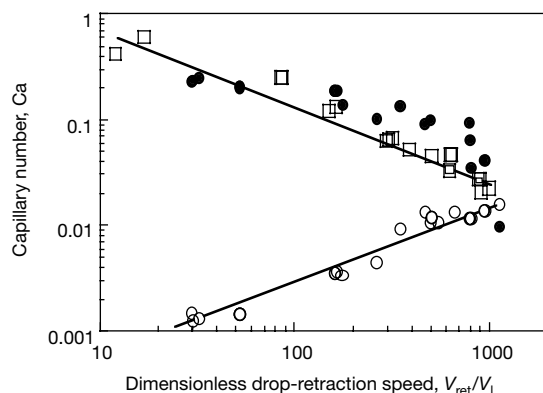


Figure 3 Capillary number versus non-dimensional drop-retraction speed, (V_{ret}/V_l), of the droplets; squares represent water-glycerol mixtures and circles represent polymer solutions of different concentrations. Open symbols denote that the capillary number has been calculated using the shear viscosity: a strong disagreement with the newtonian data is found in this case, as highlighted by the solid lines. In contrast, the agreement is very good when the capillary number is calculated using the elongational viscosity (filled circles).

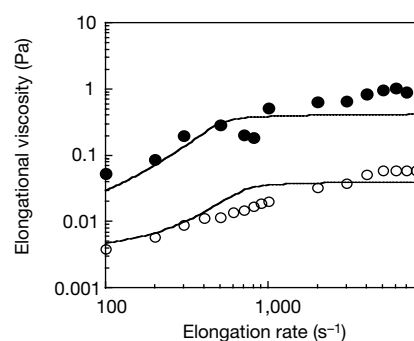


Figure 4 Elongational viscosity versus elongation rate. Measurements made with a Rheometrics RFX opposing nozzle for two dilute PEO polymer solutions are shown: 0.25 g l⁻¹, molecular weight 4×10^6 (filled circles); 1.0 g l⁻¹, molecular weight 2×10^6 (open circles). Lines represent the corresponding FENE-P model calculations. A Rheologica Stress-Tech rheometer equipped with a cone-plate geometry was used to obtain the data needed for our FENE model calculations. The cone is 55 mm in diameter and has an angle of 0.5°, which allows the measurement of extremely small normal forces at very high shear rates. We have experimentally determined by viscosity measurements that the overlap concentration, c^* , is about 0.5 g l⁻¹ for our polymer solutions. This indicates a concentration of 0.5 g l⁻¹ as the experimentally determined value for the dilute regime and that nearly all of the experiments shown fall within the so-called 'dilute regime' for the conditions of our study.

Evidence that the elongational viscosity is the predominant factor influencing the difference in drop-retraction behaviour between the newtonian and non-newtonian solutions can be demonstrated by reformulating the capillary number using the elongational, rather than the shear viscosity. Thus, we define the capillary number for our retracting drops as $Ca_e = V_{ret}\eta_e/3\sigma$, noting that the elongational viscosity for newtonian liquids is given by $\eta_e = 3\eta$. If the capillary number is recalculated in this way, the results for the dilute polymer solutions coincide with those of the newtonian solutions (Fig. 3). Therefore, we see that here it is the elongational viscosity that dominates during drop retraction, consequently inhibiting droplet rebound.

Finally, we note that our observations are general and that we have observed the same phenomena for different hydrophobic surfaces: polished Teflon and paper hydrophobized by treatment with polydimethylsilicone (for example, 'Silkraf 9564' from Ahlstrom). In addition, other dilute solutions of high molecular weight flexible polymers, such as locus-bean and guar gum, manifest high elongational viscosities and significantly reduce droplet rebound¹⁴. Significant improvements in drop deposition on rough surfaces has also been witnessed, however, rigorous quantification of these effects remains the subject of future work.

In conclusion, we have shown that drop rebound, a major problem in a number of industrial applications, can be inhibited very effectively by the addition of small amounts of a flexible polymer. We have also shown that the critical property provided by these polymers is a high elongational viscosity. This non-newtonian property dampens the drops' retraction after impact, which in turn prevents droplet rebound. We propose and validate a simple way of evaluating the elongational viscosity of dilute solutions of flexible polymers. The solutions can be very dilute, so that the shear viscosity is essentially the same as that of the solvent. This last point has important practical implications, as drop impact will nearly always be preceded by other processes (for example, pumping) for which a low shear viscosity is needed. □

Received 30 September 1999; accepted 27 April 2000.

- Wirth, W., Storp, S. & Jacobsen, W. Mechanisms controlling leaf retention of agricultural spray solutions. *Pest. Sci.* **33**, 411–422 (1991).
- Chandra, S. & Avedisian, C. T. On the collision of a droplet with a solid surface. *Proc. R. Soc. Lond. A* **432**, 13–41 (1991).

3. Rein, M. Phenomena of liquid drop impact on solid and liquid surfaces. *Fluid Dynamics Res.* **12**, 61–93 (1993).
4. Fukai, J. *et al.* Wetting effects on the spreading of a liquid droplet colliding with a flat surface: experiment and modeling. *Phys. Fluids* **7**, 236–247 (1995).
5. Bennett, T. & Poulikakos, D. Splat–quench solidification: estimating the maximum spreading of a droplet impacting a solid surface. *J. Mater. Sci.* **28**, 963–970 (1993).
6. Scheller, B. L. & Bousfield, D. W. Newtonian drop impact with a solid surface. *Am. Inst. Chem. Eng. J.* **41**, 1357–1367 (1995).
7. Mao, T., Kuhn, D. & Tran, H. Spread and rebound of liquid droplets upon impact on flat surfaces. *Am. Inst. Chem. Eng. J.* **43**, 2169–2179, (1997).
8. de Gennes, P. G. Wetting: statics and dynamics. *Rev. Mod. Phys.* **57**, 827–863 (1985).
9. Hayes, R. A. & Ralston, J. Forced liquid movement on low energy surfaces. *J. Colloid Interface Sci.* **159**, 429–438 (1993).
10. Hudson, N., & Jones, T. The A1 project—an overview. *J. Non-Newtonian Fluid Mech.* **46**, 69–88 (1993).
11. Dontula, P., Pasquali, M., Scriven, L. E. & Macosko, C. W. Can extensional viscosity be measured with opposed-nozzle devices? *Rheol. Acta* **36**, 429–448 (1997).
12. Bird, R. B., Armstrong, R. C. & Hassager, O. in *Dynamics of Polymeric Liquids* Vols 1, 2, pp 495–520 (Wiley, New York, 1987).
13. Tanner, R. I. in *Engineering Rheology* 168–224 (Clarendon, Oxford, 1992).
14. Bergeron, V., Martin, J. -Y. & Vovelle, L. Utilisation de polymères comme agents anti-rebond dans des formulations mises en œuvre en milieux aqueux. French Patent Application 9810471 (1998); International Extension PCT/FR99/02002 (1999).
15. Yerushalmi-Rozen, R., Klein, J. & Fetters, L. J. Suppression of rupture in thin, nonwetting liquid films. *Science* **263**, 793–795 (1994).
16. Yerushalmi-Rozen, R. & Klein, J. Stabilization of non-wetting thin liquid films on a solid substrate by polymeric additives. *Langmuir* **11**, 2806–2814 (1995).

Acknowledgements

D.B. thanks S. Kumar for helpful discussions; and J. Vermant and P. Moldenaers for help with the elongational viscosity measurements. LPS de l'ENS is UMR 8550 of the CNRS, associated with the universities Paris 6 and 7. We also thank the European Commission for their support in the form of TMR project funding.

Correspondence and requests for materials should be addressed to V.B.
(e-mail: vance.bergeron@rhone-poulenc.com).

Significant dissipation of tidal energy in the deep ocean inferred from satellite altimeter data

G. D. Egbert* & R. D. Ray†

* College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, Oregon 97331, USA

† NASA Goddard Space Flight Center, Code 926, Greenbelt, Maryland 20771, USA

How and where the ocean tides dissipate their energy are long-standing questions¹ that have consequences ranging from the history of the Moon² to the mixing of the oceans³. Historically, the principal sink of tidal energy has been thought to be bottom friction in shallow seas^{4,5}. There has long been suggestive evidence^{6,7}, however, that tidal dissipation also occurs in the open ocean through the scattering by ocean-bottom topography of surface tides into internal waves, but estimates of the magnitude of this possible sink have varied widely^{3,8–11}. Here we use satellite altimeter data from Topex/Poseidon to map empirically the tidal energy dissipation. We show that approximately 10¹² watts—that is, 1 TW, representing 25–30% of the total dissipation—occurs in the deep ocean, generally near areas of rough topography. Of the estimated 2 TW of mixing energy required to maintain the large-scale thermohaline circulation of the ocean¹², one-half could therefore be provided by the tides, with the other half coming from action¹³ on the surface of the ocean.

Vertical mixing rates in the deep ocean implied by ocean microstructure¹⁴ and tracer-release data¹⁵ are typically an order of magnitude too small to balance the rate at which dense bottom

water is created at high latitudes¹². It has thus been suggested that much of the mixing required to maintain the abyssal stratification, and hence the large-scale meridional overturning, occurs at localized 'hotspots' near areas of rough topography^{4,16,17}. Numerical modelling studies further suggest that the ocean circulation is sensitive to the spatial distribution of vertical mixing¹⁸. Thus, clarifying the physical mechanisms responsible for this mixing is important, both for numerical ocean modelling and for general understanding of how the ocean works. One significant energy source for mixing may be barotropic tidal currents. The likelihood of this depends *inter alia* on whether enough power is being extracted from the barotropic tides in the likely 'hotspot' regions. Topex/Poseidon (T/P) satellite altimeter data have now made possible accurate mapping of open-ocean tidal elevations¹⁹, and provide a new opportunity to quantify empirically the spatial localization of tidal dissipation.

Here we concentrate on the principal lunar semi-diurnal tide M₂, which accounts for approximately two-thirds of the total planetary dissipation²⁰, and is the most accurately known tide. This accuracy is needed, because mapping dissipation involves second-order gradients of measured fields and small differences of large numbers.

The barotropic ocean tide satisfies the tidal equations of Laplace, modified to include effects of an elastic Earth and a self-gravitating ocean²¹. These equations embody conservation of momentum and mass for the ocean fluid:

$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{f} \times \mathbf{U} = -gH\nabla(\zeta_{\text{EQ}} - \zeta_{\text{SAL}}) - F \quad (1)$$

$$\frac{\partial \zeta}{\partial t} = -\nabla \cdot \mathbf{U} \quad (2)$$

Here ζ is the tidal elevation; $\mathbf{U}$ is the volume transport vector, equal to velocity times water depth H ; $\mathbf{f}$ is the Coriolis parameter (oriented to the local vertical), and F is a generic frictional or dissipative stress. The forcing is specified through an equilibrium tide ζ_{EQ} , which must allow for the Earth's body tide²¹, and an equilibrium-like 'tide' ζ_{SAL} which is induced by the tide's self-attraction and loading²².

Equations (1) and (2) may be combined and averaged over time to obtain an expression for the local balance between work rate, W , energy flux, P , and dissipation rate, D :

$$W - \nabla \cdot \mathbf{P} = D \quad (3)$$

A number of different explicit forms for this balance appear in the literature^{21,23}, reflecting various groupings of terms, and different definitions for work and flux (as well as omission of supposed secondary terms like self-attraction). Here we adopt simple expressions for these terms based directly on equations (1) and (2):

$$\mathbf{P} = \rho g \langle \mathbf{U} \zeta \rangle \quad W = \rho g \langle \mathbf{U} \cdot \nabla (\zeta_{\text{EQ}} + \zeta_{\text{SAL}}) \rangle \quad (4)$$

where the brackets $\langle \rangle$ denote time averages, ρ is mean seawater density and g is gravitational acceleration. Note that W represents the mean rate of working on the ocean of all tidal gravitational forces (including self-attraction forces) and of the moving ocean bottom.

Topex/Poseidon altimeter data provide a direct constraint on tidal elevations ζ , and a number of nearly global maps are now

Table 1 Partition of M₂ energy dissipation (in terawatts) between shallow seas and the deep ocean

	TPXO.4a	GOT99hf	TPXO.4b	TPXO.4c	GOT99nf	Prior	Error
Shallow seas	1.60	1.71	1.61	1.62	1.87	1.96	0.06
Deep ocean	0.83	0.74	0.82	0.81	0.57	0.06	0.06
Total	2.44	2.45	2.43	2.43	2.44	2.02	0.01

Results (in TW) are presented for the five empirical estimates discussed in the text, and for the purely hydrodynamic solution used as the prior model for the assimilation. Error bars²⁸ are for the assimilation solution TPXO.4a. Here shallow seas are defined to include all ocean areas landward of the thin line in Fig. 1b.

available. Comparison to pelagic and island tide gauges suggest¹⁹ accuracies in the open ocean of roughly 1 cm. Given ζ , calculation of ζ_{SAL} is straightforward²², while the equilibrium tide ζ_{EQ} is of course known to high accuracy. The primary challenge in using equation (3) to calculate D thus lies in finding the volume transports required for equation (4). We estimate $\mathbf{U}$ by fitting equations (1) and (2), along with no-flow boundary conditions at the coasts, in a weighted least-squares sense. Note that the fit cannot in general be exact, because with ζ given the tidal equations are overdetermined. This calculation requires specifying the form of the frictional stress F , so it is crucial to show that the resulting D from equation (3) is insensitive to this assumption. We do this by using two separate tidal elevation solutions based on two distinct approaches, and by varying relevant parameters over a wide range. Five of the resulting dissipation estimates are used below; a more comprehensive set will be published elsewhere.

The key to the stability of these calculations is a proper choice of weights for the least-squares fitting problem. Because equation (2) is just a simple statement of mass conservation and does not depend on any unknown parameters, this equation should be fitted to a precision consistent with the T/P measurements of ζ —to about 1 cm. Larger relative misfits are justified for equation (1) to allow for uncertainties in bathymetry and in dissipation. By relaxing the fit to the momentum conservation equations, we allow the empirically determined T/P elevations to make appropriate adjustments to the dynamical balance. In particular, residuals to equation (1) can formally do work²⁴ to enhance or reduce dissipation where our *a priori* assumptions about friction are inconsistent with the observed ζ .

Transports could also be estimated by simply substituting ζ into equation (1) and solving for $\mathbf{U}$. But with this approach the fit to equation (1) is exact and all misfit must be accommodated in equation (2). This results in transports that do not conserve mass, and in dissipation estimates that are too noisy to be useful.

The two elevation solutions considered here are denoted TPXO.4 and GOT99, both based on about six years of T/P measurements. For the TPXO.4 solution, we used a variational data assimilation scheme in which both ζ and $\mathbf{U}$ were estimated by directly fitting T/P altimeter data to the tidal equations^{24,25}. This requires a prior model for which we used a purely hydrodynamic (time-step) solution with

quadratic bottom friction. The fit to equations (1) and (2) was controlled by a non-local, and spatially variable, dynamical error covariance, of three very different forms: for TPXO.4a, the covariances were as described in ref. 25; for TPXO.4b, a spatially uniform covariance was assumed; and for TPXO.4c, dynamical error variances were increased in regions of rough topography.

For the GOT99 solutions, gridded elevation fields²⁶ ζ were substituted into equation (1) and (2), and $\mathbf{U}$ was estimated by weighted least-squares²⁷, consistent with the errors in ζ . Bottom stress was parametrized as $F = r\mathbf{U}/H$, with r ranging from 0.0 (no frictional dissipation at all, denoted GOT99nf) to 0.03 m s^{-1} (an inordinately large friction, denoted GOT99hf).

The deduced dissipation estimates TPXO.4a and GOT99hf are plotted in Fig. 1a and 1b, respectively. Note that the colour scale has been chosen to emphasize the pattern of dissipation in the open ocean, and that the scale saturates in some shallow seas where dissipation is generally greatest. In both plots there are some blue patches where the estimated dissipation is negative. This is clearly not physical, and provides some indication of the magnitude of noise in the estimates. Results from the assimilation method are somewhat smoother and overall have a cleaner appearance (that is, fewer blue spots).

Although there are many differences in detail, the two dissipation estimates have many features in common. The areas of intense dissipation expected due to bottom boundary layer drag in shallow seas are clearly evident in both plots (for example, in Hudson Bay and the Yellow Sea; on the European, Patagonian and northwest Australia continental shelves). More interestingly, the estimates exhibit a similar pattern of significant dissipation in the open ocean. For example, dissipation is clearly enhanced in the Pacific Ocean over the Hawaiian ridge, the Tuamotu archipelago, and over the back arc island chains extending from Japan southwards to New Zealand. The western Indian Ocean around the Mascarene ridge and south of Madagascar also exhibits enhanced dissipation in both estimates. The Mid-Atlantic Ridge shows up clearly in the TPXO.4a assimilation estimate (Fig. 1a), especially in the south Atlantic Ocean. Dissipation is also enhanced in this area in the GOT99hf estimates (Fig. 1b), though the pattern is less distinctive.

All of the areas where dissipation is consistently enhanced are characterized by significant bathymetric roughness, generally with

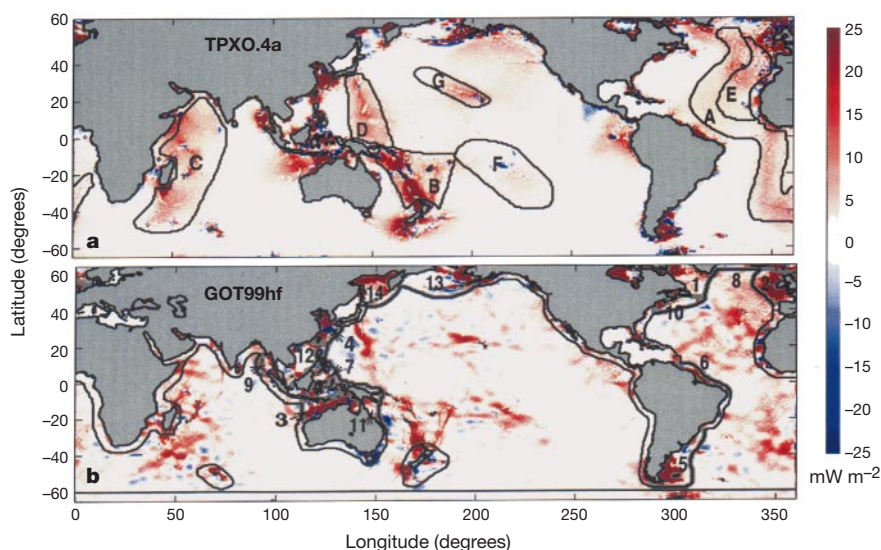


Figure 1 Estimates of M_2 tidal energy dissipation. These estimates of D are computed for two recent Topex/Poseidon (T/P) M_2 tidal solutions. **a**, TPXO.4a elevations and transports, estimated using a variational data assimilation method²⁵. **b**, GOT99hf with transports obtained by least-squares fitting²⁷ of equations (1) and (2) to gridded elevations estimated

from the altimetry data²⁶. Deep and shallow ocean areas discussed in the text and in Fig. 2 are outlined and labelled in **a** (deep) and **b** (shallow). The thin line in **b** is the boundary between deep ocean and shallow seas.

elongated features such as ridges and island chains oriented perpendicular to tidal flows.

To quantify the large-scale pattern of tidal energy dissipation we consider integrals of D over some larger ocean areas, as outlined in Fig. 1. These integrals can be expressed as a sum of work terms involving integrals of tidal elevations over the whole patch, and line integrals of flux through the boundary of the patch²⁴. Only the latter require estimates of transports, so by placing boundaries in deep water the need to estimate currents in shallow seas can be avoided. Total dissipation in shallow seas may thus be accurately determined, even when details of currents in these areas are poorly resolved by the altimeter data.

In Fig. 2a we plot total dissipation for all shelf and shallow-sea areas that account for at least 25 GW of dissipation each. The areas used for these calculations are outlined by heavy solid lines and numbered in Fig. 1b. Results are plotted for all five dissipation estimates—three based on the assimilation solutions, and two based on the weighted least-squares fit to the GOT99 elevations. Also plotted are estimates of error bars for the TPXO.4a dissipation estimates, computed using the Monte Carlo approach described in ref. 28. For most of these major shallow-sea sinks, all dissipation estimates agree within approximately 10–15%; an exception is the complex seas around Indonesia where neither currents nor elevations are well mapped. Note that the region denoted Hudson Bay also includes the Labrador Sea, Baffin Bay and all points north and west.

We also calculated total dissipation in the seven open ocean areas indicated in Fig. 1a as A–G. These are plotted in Fig. 2b. The most significant areas of open ocean dissipation include the Mid-Atlantic Ridge, the western Pacific around Fiji, and the western Indian Ocean. Each of these accounts for approximately 100 GW, comparable to major shallow-sea sinks.

In Table 1 we summarize the partition of dissipation between shallow and deep ocean areas (as defined by the light solid line in

Fig. 1b) for the five dissipation estimates of Fig. 2. For comparison, we also include in Table 1 dissipation totals for the purely hydrodynamic prior solution used for all of the assimilation solutions, along with posterior error bars²⁸ for TPXO.4a. The estimates of global M_2 total dissipation, which in fact depend only on the tidal elevations^{21,20}, converge to about 2.44 TW, consistent with independent estimates based on space-geodetic techniques, after allowance is made for solid-earth and atmospheric dissipation²⁹. The prior solution, which is not constrained by the T/P data, dissipates only 2.01 TW, consistent with results reported for other purely hydrodynamic tidal solutions²³. Its dissipation occurs only in shallow seas where tidal current speeds are by far the greatest. In contrast, for all of the T/P-constrained solutions a substantial fraction of the dissipation (0.6–0.8 TW out of 2.4 TW, or 25–30%) occurs in the deep ocean.

The TPXO.4a error bars for deep ocean dissipation (0.06 TW) do not cover the full range of estimates in Table 1, and thus do not adequately describe our present uncertainty about the magnitude of tidal dissipation in the deep ocean. The statistical model used for the error calculation was ‘tuned’ for consistency with observed errors in the data and dynamics²⁸, and it probably accounts reasonably well for the effects of noise and for limitations in the altimeter data coverage. However, these error bars cannot account for systematic biases which could be implicit in some (or all) of the different estimation schemes. This could particularly be an issue for the GOT99 solutions, which purposely employed a range of frictional drag coefficients far exceeding physically plausible values. Although Table 1 suggests that assumptions about dynamics may have some effect on the final dissipation estimates, similar dissipation totals and a similar large-scale dissipation pattern over the globe are in fact obtained in all cases. Taking account of the formal error bars and the variations between estimates calculated under different assumptions, we estimate that deep sea dissipation for M_2 is approximately 0.7 ± 0.15 TW. The enlarged error bars are probably pessimistic,

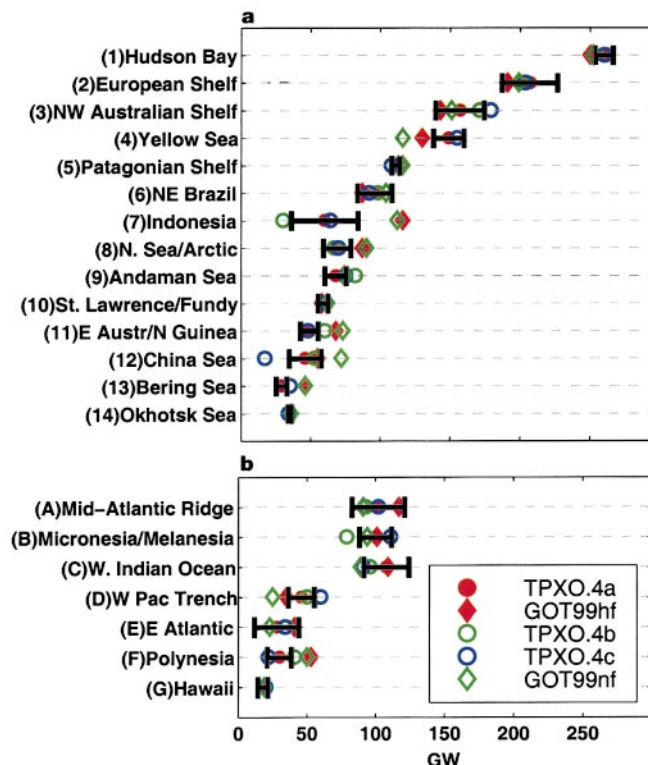


Figure 2 Area-integrated dissipation for selected shallow seas and deep-ocean areas. Results are presented for the five estimates discussed in the text, along with posterior error bars²⁸ for the assimilation solution TPXO.4a. **a**, Principal shallow-sea sinks of tidal

energy. Locations and boundaries of the numbered areas are given in Fig. 1b. **b**, Dissipation in selected open ocean areas, labelled A–G in Fig. 1a.

with the lower limit of 0.55 TW (which is approached only when currents are computed under the unrealistic assumption of no friction) being particularly unlikely.

Although our results do not explicitly constrain the mechanism of open-ocean tidal dissipation, the observed spatial pattern strongly suggests a significant role for scattering into internal tides, induced by tidal flow of the stratified ocean over rough topography^{8,10,11}. Indeed, many of the open-ocean areas of high dissipation are known generators of strong internal tides. For example, internal tide observations along the Hawaiian ridge have suggested³⁰ an energy flux of roughly 15 GW propagating away from the ridge, which is not far from our 20 GW estimate for Hawaii (Fig. 2b). The T/P dissipation estimates are also qualitatively consistent with previous simplified theoretical models^{8,11} of internal-tide energetics, although our empirical results suggest that these previous model estimates are somewhat too large.

Extrapolating the results for M_2 to all lunar and solar constituents suggests that approximately 1 ± 0.25 TW of tidal power is dissipated in the deep ocean, probably by scattering into internal waves. This process may also be at work to some extent along the edges of continental shelves (which lie entirely within the area we have classified as shallow sea), so our global estimate is likely to be conservative.

In a recent discussion of the ocean's thermohaline circulation (or, more strictly, its meridional overturning), Munk and Wunsch¹² estimate that 2 TW is needed to maintain the global abyssal density distribution, and they list the possible mechanisms that might provide the required power. They note that the wind provides 1 TW, and ask whether the ocean tides can provide the other. Our answer is that the tides can provide it. But there remain many questions: about the process of barotropic/baroclinic conversion, about the ultimate decay of internal tides into turbulence, and about the implications of these processes for large-scale ocean circulation and climate. □

Received 9 November 1999; accepted 5 April 2000.

- Munk, W. H. & MacDonald, G. J. F. *The Rotation of the Earth: A Geophysical Discussion* (Cambridge Univ. Press, 1960).
- Hansen, K. S. Secular effects of oceanic tidal dissipation on the moon's orbit and the earth's rotation. *Rev. Geophys. Space Phys.* **20**, 457–480 (1982).
- Munk, W. H. Once again: once again—tidal friction. *Prog. Oceanogr.* **40**, 7–36 (1997).
- Taylor, G. I. Tidal friction in the Irish Sea. *Phil. Trans. R. Soc. Lond. A* **220**, 1–33 (1919).
- Jeffreys, H. Tidal friction in shallow seas. *Phil. Trans. R. Soc. Lond. A* **221**, 239–264 (1920).
- Munk, W. H. Abyssal recipes. *Deep-Sea Res.* **13**, 707–730 (1966).
- Bell, T. H. Topographically induced internal waves in the open ocean. *J. Geophys. Res.* **80**, 320–327 (1975).
- Morozov, E. G. Semidiurnal internal wave global field. *Deep-Sea Res.* **42**, 135–148 (1995).
- Wunsch, C. Internal tides in the ocean. *Rev. Geophys. Space Phys.* **13**, 167–182 (1975).
- Baines, P. G. On internal tide generation models. *Deep-Sea Res.* **29**, 307–338 (1982).
- Sjöberg, B. & Stigebrandt, A. Computations of the geographical distribution of the energy flux to mixing processes via internal tides and the associated vertical circulation in the ocean. *Deep-Sea Res.* **39**, 269–291 (1992).
- Munk, W. H. & Wunsch, C. Abyssal recipes II: Energetics of tidal and wind mixing. *Deep-Sea Res.* **45**, 1977–2010 (1998).
- Wunsch, C. The work done by the wind on the ocean circulation. *J. Phys. Oceanogr.* **28**, 2331–2339 (1998).
- Gregg, M. C. Scaling turbulent dissipation in the thermocline. *J. Geophys. Res.* **94**, 9686–9698 (1989).
- Ledwell, J. R., Watson, A. J. & Law, C. S. Evidence for slow mixing across the pycnocline from an open ocean tracer release experiment. *Nature* **364**, 701–703 (1993).
- Armi, L. Some evidence for boundary mixing in the deep sea. *J. Geophys. Res.* **83**, 1971–1979 (1978).
- Polzin, K. L., Toole, J. M., Ledwell, J. R. & Schmitt, R. W. Spatial variability of turbulent mixing in the abyssal ocean. *Science* **276**, 93–96 (1997).
- Samelson, R. M. Large scale circulation with locally enhanced vertical mixing. *J. Phys. Oceanogr.* **28**, 712–726 (1998).
- Shum, C. K. *et al.* Accuracy assessment of recent ocean tide models. *J. Geophys. Res.* **102**, 25173–25194 (1997).
- Cartwright, D. E. & Ray, R. D. Energetics of global ocean tides from Geosat altimetry. *J. Geophys. Res.* **96**, 16897–16912 (1991).
- Hendershott, M. C. in *The Sea* (eds Goldberg, E. *et al.*) 47–95 (Wiley, New York, 1977).
- Ray, R. D. Ocean self-attraction and loading in numerical tidal models. *Mar. Geodesy* **21**, 181–192 (1998).
- Le Provost, C. L. & Lyard, F. Energetics of the barotropic ocean tides: An estimate of bottom friction dissipation from a hydrodynamic model. *Prog. Oceanogr.* **40**, 37–52 (1997).
- Egbert, G. D. Tidal data inversion: interpolation and inference. *Prog. Oceanogr.* **40**, 81–108 (1997).
- Egbert, G. D., Bennett, A. F. & Foreman, M. G. G. TOPEX/POSEIDON tides estimated using a global inverse model. *J. Geophys. Res.* **99**, 24821–24852 (1994).

- Ray, R. D. *A Global Ocean Tide Model from TOPEX/POSEIDON Altimetry: GOT99.2* (NASA/TM-1999-209478, Goddard Space Flight Center, Greenbelt, Maryland 1999).
- Ray, R. D. Inversion of oceanic tidal currents from measured elevations. *J. Mar. Sys.* (submitted).
- Dushaw, B. D. *et al.* A TOPEX/POSEIDON global tidal model (TPXO.2) and barotropic tidal currents determined from long range acoustic transmissions. *Prog. Oceanogr.* **40**, 337–369 (1998).
- Ray, R. D., Eanes, R. J. & Chao, B. F. Detection of tidal dissipation in the solid Earth by satellite tracking and altimetry. *Nature* **381**, 595–597 (1996).
- Ray, R. D. & Mitchum, G. T. Surface manifestation of internal tides in the deep ocean: observations from altimetry and island gauges. *Prog. Oceanogr.* **40**, 135–162 (1997).

Acknowledgements

We thank W. Munk for discussions. This work was supported by the US National Science Foundation (G.D.E.) and the US National Aeronautics and Space Administration (R.D.R.).

Correspondence and requests for materials should be addressed to G.D.E. (e-mail: egbert@oce.orst.edu).

A larger pool of ozone-forming carbon compounds in urban atmospheres

Alastair C. Lewis[†], Nicola Carslaw^{*}, Philip J. Marriott[‡], Russel M. Kinghorn[‡], Paul Morrison[‡], Andrew L. Lee^{*}, Keith D. Bartle^{*} & Michael J. Pilling^{*}

^{*} School of Chemistry, [†] School of the Environment, University of Leeds, Woodhouse Lane, Leeds LS2 9JT, UK

[‡] Royal Melbourne Institute of Technology, Department of Applied Chemistry, Melbourne 3001, Victoria, Australia

Volatile organic compounds play a central role in the processes that generate both urban photochemical smog and tropospheric ozone^{1,2}. For successful and accurate prediction of these pollution episodes, identification of the dominant reactive species within the volatile organic carbon pool is needed³. At present, lack of resolution inherent in single-column chromatographic analysis⁴ limits such a detailed chemical characterization of the complex urban atmosphere. Here we present an improved method of peak deconvolution from double-column (orthogonal) gas chromatography^{5,6}. This has enabled us to isolate and classify more than 500 chemical species of volatile organic compounds in urban air, including over 100 multi-substituted monoaromatic and volatile oxygenated hydrocarbons. We suggest that previous assessments of reactive carbon species may therefore have underestimated the contribution made by volatile organic compounds to urban pollution, particularly for compounds with more than six carbon atoms. Incorporating these species in predictive models should greatly improve our understanding of photochemical ozone yields and the formation of harmful secondary organic aerosols^{7,8}.

Previous studies have shown that ozone production in urban areas of Europe is limited by the availability of volatile organic compounds (VOCs)¹. Under these conditions, the concentration of ozone increases with that of VOCs, as demonstrated by standard ozone isopleth diagrams. Methods have been developed to rank the ability of individual VOCs to produce ozone, resulting in the US MIR scale⁹ and the UK POCP scale⁷. Despite differences in definition, mode of calculation and emissions inventories, there is a broad general agreement on the relative ordering of different VOCs⁷. Generally, volatile aromatics have high reactivity (and hence POCP values), which increases with the degree of alkyl substitution. In addition to high photochemical reactivity, there is now strong evidence that the photochemistry of aromatic species can lead to the formation of secondary organic aerosols (SOA), which are known to

be harmful to human and ecosystem health⁸. As emissions of aromatic species are concentrated in urban areas, the formation of SOA may become an acute public health concern. We also note that many aromatic compounds are carcinogenic and/or mutagenic, particularly when nitrated.

Here, we show that a large proportion of volatile organic carbon is aromatic in nature, with clear consequences for ozone formation. To demonstrate this, a series of urban air samples were collected from the centre of Melbourne, Australia, and analysed using the methodology described. Figure 1 shows an orthogonally generated (comprehensive) chromatogram, alongside a concurrent one-dimensional separation. Visual inspection indicates that greater numbers of peaks are individually isolated using the comprehensive approach. It is difficult to visualize the entire chromatogram under high gain, since baseline drift intensity changes can obscure peaks in different regions of the chromatogram. Taking smaller sections of the chromatogram, enhancement in gain may be made, as shown in Fig. 2 for an urban air sample. In Fig. 2, the single-column separation has 20–30 visible peaks above the baseline, although many are clearly co-eluting, whereas the comprehensive separation shows 110–120 components. The separation demonstrates that in a high-complexity matrix, several compounds potentially co-elute at any primary column retention time, so that single-column separations in isolation must be interpreted with great care. When individual sections of this chromatogram are expanded in a manner similar to Fig. 2, a total of over 550 individual components are isolated.

Large numbers of VOCs clearly exist in the higher boiling ranges above C₆ (greater than 6 carbon atoms), isolated in ambient air using only comprehensive chromatography. In many regions of the

single-column separation, no single component is present at sufficient concentration to be elevated above what is erroneously described as the baseline. The baseline obtained from single-dimension separation is, therefore, a composite of the elution of many hundreds of low-concentration species. With a flame ionization detector this composite response is manifested as raised standing current at the collector electrode. An analogous phenomenon is observed when using equivalent full-scan gas chromatography–mass spectroscopy, indicating a constant background in basic hydrocarbon fragmentation units, usually attributed to high-vacuum leaks or chemical impurity. These species, whose response is within the 'baseline', represent significant atmospheric carbon abundance, particularly when the effects of reduced response per carbon atom is corrected for with changing structural and functional composition.

Although instruments for the analysis of bulk total organic carbon exist, reconciliation with comprehensive gas-chromatography analysis may lead to an underestimation in organic content. As the carbon number increases it becomes increasingly important that detector calibration is specific to both carbon number and functional group (notably for oxygen groupings)¹⁰, and this can only be possible when full deconvolution of the sample matrix is performed.

As only peaks substantially above the nominal baseline level may be determined, (normally at signal to noise S/N = 3 : 1), conventional chromatography clearly underestimates total concentration. Measurement of C₂–C₆ hydrocarbon species using a single-column method (that of Lewis *et al.*¹¹) for which resolution is complete, coupled with either single column or comprehensive analysis for C₆–C₁₄, allows an assessment of total volatile organic carbon mass over the entire C₂–C₁₄ range, by determining all area counts from the distinguishable peaks above baseline. Single-column methods indicate that approximately 60% of total reactive carbon mass lies within the C₂–C₆ range and approximately 40% in the range C₆–C₁₄. Repeating this determination using C₆–C₁₄ values calculated from comprehensive gas chromatography, the fraction of carbon in the C₂–C₆ range falls to less than 20% of total volatile carbon. Thus conventional chromatography fails to determine around two-thirds of the carbon mass.

The selectivity of the secondary column allows broad chemical classifications to be made of the sample as a whole. Figures 1 and 2

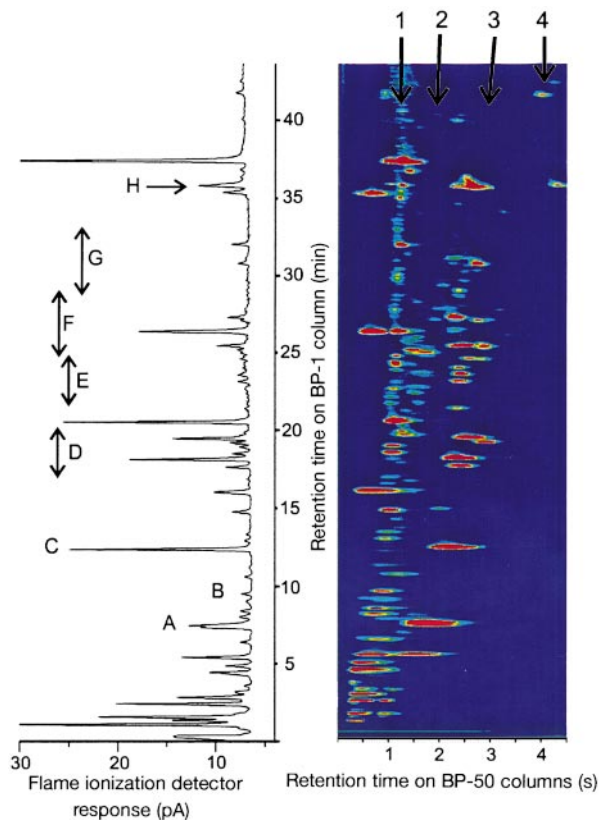


Figure 1 Comprehensive and one-dimensional separations of volatile organic compounds in urban air. A, benzene; B, heptane; C, toluene; D, xylenes; E, C₃-benzenes; F, C₄-benzenes; G, C₅-benzenes; H, naphthalene. 1, aliphatic band; 2, carbonyl band; 3, aromatic band; 4, bi-aromatics.

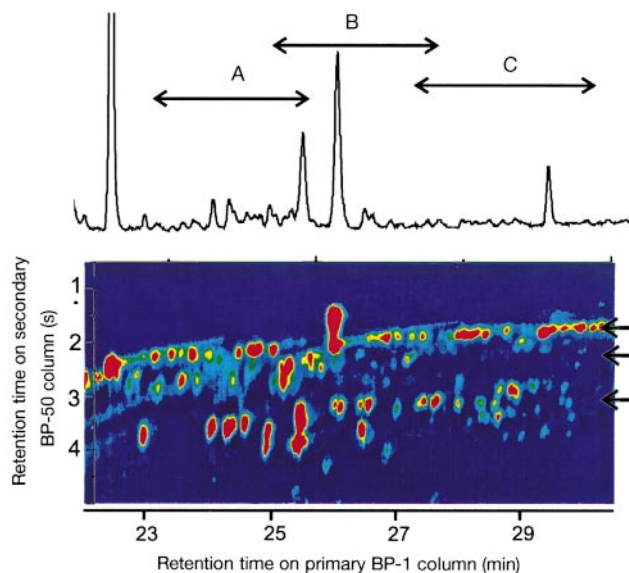


Figure 2 Expanded section of volatile organic compounds in an urban air separation. Upper trace, single-dimension separation with the flame ionization detector. Lower trace, comprehensive separation with the flame ionization detector. A, C₃-benzenes; B, C₄-benzenes; C, C₅-benzenes. 1, aliphatic band; 2, carbonyl band; 3, aromatic band.

are marked with distinct bands of compounds that can be described crudely as aromatic, aliphatic or carbonyl. The observations indicate that aliphatic and aromatic species have a relatively constant profile, whereas carbonyl species vary considerably depending on time of day and meteorological conditions. A substantial amount of chemically stable, physically volatile, oxygenated material is isolated in the C_9 – C_{11} region, comprising around 85% of all oxygenated material.

A summation has been made of the total area above baseline for the three chemical classifications indicated in Fig. 2. The highest fractional contribution of oxygenated species was observed to occur during the period 14:30 to 17:00, consistent with formation from photochemical processes rather than primary emission. Distributions of C_6 – C_{14} organics during this time period were: total aliphatic, $45.2 \pm 1.6\%$; total aromatic, $51.1 \pm 1.9\%$; and total oxygenated, $2.0 \pm 0.6\%$.

The calculated ratio of aromatic to alkane species in this study is significantly higher than has been reported previously using single-column methodologies. We believe this observation is attributable to the significantly greater numbers of aromatic species isolated and subsequently included in the comprehensive gas chromatography aromatic summation, and that this ratio is not a result of a unique emission source. This study has focused on only a single city, so we must examine localized emission sources to establish the degree to which this underestimation may be occurring in other urban areas world-wide.

A comparison of comprehensive separations of urban air with those of gasoline and diesel vapours shows very strong similarities in both composition and relative distributions, suggesting that these are the most significant sources of increased aromatic loading. In general terms, three-dimensional data sets from comprehensive separations offer emission-source deconvolution using pattern recognition and matching technology, and examples of this application have recently been reported in the area of oil-spill fingerprinting¹².

Emissions inventories for the Melbourne area are broadly similar in both source type and chemical composition to those reported for

the United Kingdom¹³, suggesting that the observations are applicable to urban areas of other industrialized nations. In almost all national inventories mobile and evaporative sources are poorly described, because of highly variable fuel composition for even a single location. It has been reported that the contribution of evaporative emissions (both $<C_6$ and $>C_6$) from mobile sources to the total hydrocarbon burden is very severely underestimated by inventory data in both US and Australian cities¹⁴, and our observations appear to support this hypothesis. In much of continental Europe, diesel fuel usage is increasing, shifting the balance of emissions in favour of longer-chain aliphatic and aromatic components. In such locations, therefore, the degree of underestimation in volatile carbon budget may be even greater than that reported for the Melbourne area.

The observation of increased reactive carbon loading is of importance given reasonable agreement between measured ozone concentrations and those predicted by models constrained by current atmospheric measurement of VOC input. Given that extra organic carbon is often not required to balance modelled photo-oxidant budgets, flaws may exist in the way these models are parametrized or constrained. For such predictive ozone models, explanations may lie in an incomplete understanding and treatment of both gas phase and heterogeneous carbon cycling in the polluted atmosphere.

The atmospheric modelling of VOCs is generally performed using some degree of lumping of chemical mechanisms, enabling models to be of tractable size and complexity^{15,16}. To fully simulate the effects of large numbers of species greater than C_6 will therefore require model inclusion at an appropriate level of mechanism complexity in conjunction with accurate assessments of emissions. At present, however, many currently uncertain reaction pathways further complicate the generation of reduced schemes for many substituted mono-aromatics.

Describing organic carbon in terms of increased complexity may ultimately produce only small differences to calculated ozone concentrations when compared to simple VOC representations. What is vital, however, is that comprehensive quantitative assessments

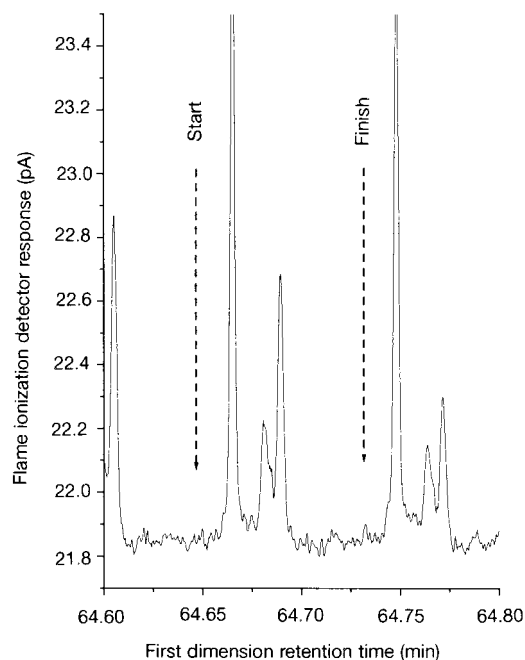


Figure 3 Fast second-dimension chromatograms generated from midpoint zone compression in a comprehensive gas chromatography separation. Peak widths at half-height are about 100 ms, around 10 to 100 times narrower than encountered in single-column gas chromatography.

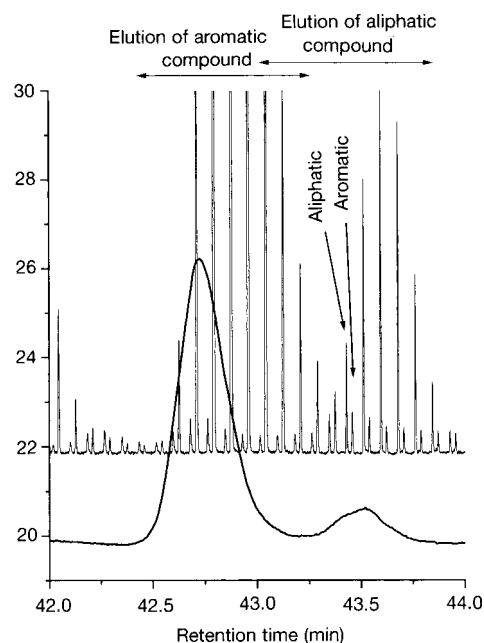


Figure 4 Comparison of column eluent from concurrent one- and two-dimensional gas chromatography separations monitored by parallel flame ionization detectors. Upper trace, output from the flame ionization detector following comprehensive separation. Lower trace, output from the flame ionization detector following single-column separation.

of both emission and atmospheric burden are used as starting points for modelling studies. It is clear that for the prediction of ozone values derived from emission inventory VOC input, it is of great importance that accurate representation of the magnitude of sources is included. One of the main conclusions of the 1997 PORC report¹ was that hydrocarbon emission inventories are biased low, in agreement with the measurements presented here.

Additional reactive carbon also has significant implications for the prediction of tropospheric OH abundance and oxidative capacity. In field studies where modelled concentrations of OH have been compared with measurements, the modelled values have tended to over-predict measurement (ref. 17 and references therein). In many cases this discrepancy has been attributed to reactions with unmeasured hydrocarbon species¹⁸, although previous analyses have been unable to substantiate this theory. Comprehensive analysis offers direct evidence in support of this hypothesis, indicating that it is volatile carbon in the range C₆–C₁₄ which is responsible, but it is currently either excluded from or incorrectly speciated in predictive model calculations. □

Experimental methods

A large-volume sample reconcentration technique was applied before comprehensive (orthogonal coupled) gas chromatography analysis¹⁹, with thermal desorption transfer from injector to primary column performed splitless. A non-polar, 50 m long, 100% methyl polysiloxane column with 0.53 mm i.d. and 5 µm film thickness was used for the primary separation, having both a large sample capacity and a strong degree of phase ratio refocusing upon injection. The secondary column was a moderately polar 50% phenyl-polysilphenylene siloxane BPX50 column, 2.2 m long, with 150 µm i.d. and 0.2 µm film thickness. The two columns were interfaced using a cryogenically modulating midpoint^{20,21} containing a 15-cm length of intermediate column, BPX50, with 150 µm i.d. and 0.5 µm film thickness. The cryogenic piston cooled 2.5 cm of intermediate column using a pneumatically regulated supply of CO₂. Each 4.5-s zone of column eluent was refocused inside the capillary column on the cold trap before pulsed introduction to column 2 by withdrawing the trap by 4 cm along the longitudinal axis of the column for a period of 0.5 s using a high-speed pneumatic drive. The modulation procedure was repeated continuously. Thermal desorption from the focused zone was extremely fast, owing to a combination of extremely low thermal mass of column, and the turbulently stirred warm air within the chromatograph.

Separation on the second column was completed within the period of intermediate trap modulation with data capture at 100 Hz to adequately characterize the eluting peaks from the second dimension. An example of a short high-speed second-dimension chromatogram is shown in Fig. 3. Typical peak widths at half height were of the order of 100–150 ms. A split union was placed before the modulator, allowing both modulated and unmodulated column output to be detected on two flame ionization detectors simultaneously. Figure 4 illustrates a section of chromatogram, with overlaid outputs from two parallel detectors. The figure demonstrates the elution of an alkane and an aromatic species, under both one- and two-dimensional separation conditions. Aliphatic species elute before the aromatic on the second column, a result of the selectivity of the polar BPX50 column used here. The one-dimensional time-response data stream obtained from the modulator and second-column configuration may be further normalized and transposed to generate a two-dimensional surface plot, with detector response as the z-axis.

Received 20 September 1999; accepted 20 April 2000.

1. Fowler, D. *et al.* Fourth Report of the Photochemical Oxidants Review Group 75–104 (United Kingdom Department of the Environment, Transport and the Regions, London, 1997).
2. Madronich, S. & Calvert, J. G. Permutation reactions of organic peroxy-radicals in the troposphere. *J. Geophys. Res.* **95**, 5697 (1990).
3. Houweling, S., Dentener, F. & Lelieveld, J. The impact of non-methane hydrocarbons on tropospheric photochemistry. *J. Geophys. Res.* **103**, 10673–10696 (1998).
4. Giddings, J. C. in *Multidimensional Chromatography: Techniques and Applications* (ed. Cortes, H. J.) 1–27 (Marcel Dekker, New York, 1990).
5. Bushey, M. M. & Jorgenson, J. W. Automated instrumentation for comprehensive two dimensional high performance liquid chromatography—capillary zone electrophoresis. *Anal. Chem.* **62**, 978–984 (1990).
6. Liu, Z. & Phillips, J. B. Comprehensive two-dimensional gas-chromatography using an on-column thermal modulator interface. *J. Chromatogr. Sci.* **29**, 227–231 (1991).
7. Derwent, R. G., Jenkin, M. E., Saunders, S. M. & Pilling, M. J. Photochemical ozone creation potentials for organic compounds in northwest Europe calculated with a master chemical mechanism. *Atmos. Environ.* **32**, 2429–2441 (1998).
8. Forstner, H. L., Flagan, R. C. & Seinfeld, J. H. Secondary aerosol from the photo-oxidation of aromatic hydrocarbons: molecular composition. *Environ. Sci. Technol.* **21**, 1345–1358 (1997).
9. Carter, W. P. L. Development of ozone reactivity scales for volatile organic compounds. *J. Air Waste Manag. Assoc.* **44**, 881–899 (1994).
10. Lewis, A. C. *et al.* Atmospheric monitoring of volatile organic compounds using programmed temperature vaporisation injection. *J. High Res. Chromatogr.* **19**, 686–690 (1996).
11. McMinn, D. G. & Hill, H. H. in *Detectors for Capillary Chromatography* (eds Hill, H. H. & McMinn, D. G.) 15–17 (Wiley, New York, 1992).

12. Gaines, R. B., Frysinger, G. S., Hendrick Smith, M. S. & Stuart, J. D. Oil spill source identification by comprehensive two-dimensional gas chromatography. *Environ. Sci. Technol.* **33**, 2106–2112 (1999).
13. *Air Emissions Inventory Port Phillip Region 4–25* (Australian Environmental Protection Authority, Melbourne, 1998).
14. Duffy, B. L., Nelson, P. F., Ye, Y., Weeks, I. A. & Galbally, I. E. Emissions of benzene, toluene, xylenes and 1,3-butadiene from a representative portion of the Australian car fleet. *Atmos. Environ.* **32**, 2693–2704 (1998).
15. Carter, W. P. L. A detailed mechanism for the gas-phase atmospheric reactions of organic compounds. *Atmos. Environ.* **24**, 481–518 (1990).
16. Stockwell, W. R., Kirchner, F., Kuhn, M. & Seinfeld, S. A new mechanism for regional atmospheric chemistry modelling. *J. Geophys. Res.* **102**, 25847–25879 (1997).
17. Carslaw, N. *et al.* Modelling OH, HO₂ and RO₂ radicals in the marine boundary layer: 1. Model construction and comparison with measurements. *J. Geophys. Res.* **104**, 30241–30255 (1999).
18. McKeen, S. A. *et al.* Photochemical modeling of hydroxyl and its relationship to other species during the tropospheric OH photochemistry experiment. *J. Geophys. Res.* **102**, 6467–6493 (1997).
19. Venkatramani, C. J. & Phillips, J. B. Comprehensive two-dimensional gas-chromatography applied to the analysis of complex-mixtures. *J. Microcolumn Separat.* **5**, 511–516 (1993).
20. Kinghorn, R. M. & Marriott, P. J. Comprehensive gas chromatography using a modulating cryogenic trap. *J. High Resol. Chromatogr.* **21**, 620–622 (1998).
21. Kinghorn, R. M. & Marriott, P. J. Modulation and manipulation of chromatographic bands by using thermal means. *Anal. Sci.* **14**, 651–660 (1998).

Acknowledgements

We thank the UK Engineering and Physical Sciences Research Council and the UK Department of the Environment, Transport and the Regions for financial support. We also acknowledge the support of Hewlett-Packard (Australia), and SGE International (Australia).

Correspondence and requests for material should be addressed to A.L. (e-mail: alyl@chem.leeds.ac.uk).

Re-Os isotopic evidence for a lower crustal origin of massif-type anorthosites

Henrik Schiellerup*, David D. Lambert†‡, Tore Prestvik*, Brian Robins§, Jannene S. McBride† & Rune B. Larsen||

* Department of Geology and Mineral Resources Engineering, Norwegian University of Science and Technology, 7491 Trondheim, Norway

† Department of Earth Sciences, PO Box 28E, Monash University, Melbourne, Victoria 3800, Australia

§ Department of Geology, University of Bergen, 5007 Bergen, Norway

|| Geological Survey of Norway, 7491 Trondheim, Norway

Massif-type anorthosites are large igneous complexes of Proterozoic age. They are almost monomineralic, representing vast accumulations of plagioclase with subordinate pyroxene or olivine and Fe–Ti oxides—the 930-Myr-old Rogaland anorthosite province in southwest Norway¹ represents one of the youngest known expressions of such magmatism. The source of the magma and geodynamic setting of massif-type anorthosites remain long-standing controversies in Precambrian geology, with no consensus existing as to the nature of the parental magmas or whether these magmas primarily originate in the Earth's mantle or crust. At present, massif-type anorthosites are believed to have crystallized from either crustally contaminated mantle-derived melts that have fractionated olivine and pyroxenes at depth² or primary aluminous gabbroic to jotunitic melts derived from the lower continental crust³. Here we report rhenium and osmium isotopic data from the Rogaland anorthosite province that strongly support a lower crustal source for the parental magmas. There is no

‡ Present address: Division of Earth Sciences, National Science Foundation, Arlington, Virginia 22230, USA.

evidence of significantly older crust in southwest Scandinavia and models invoking crustal contamination of mantle-derived magmas fail to account for the isotopic data from the Rogaland province. Initial osmium and neodymium isotopic values testify to the melting of mafic source rocks in the lower crust with an age of 1,400–1,550 Myr.

Various parental magmas have been proposed for the anorthosite suite. The common presence of megacrysts of plagioclase and aluminous orthopyroxene shows that the magma was in equilibrium with plagioclase and pyroxenes at pressures exceeding 10 kbar (refs 3–5). The constricted plagioclase stability at these pressures means that the parental melts must have been highly aluminous. Chilled margins, inferred to be representative of such parental magmas, are occasionally found in anorthosite massifs and vary from high-Al, high-Fe gabbro or norite to troctolite^{2,6}. More-evolved, jotunitic rocks are present as marginal chills to anorthosites and allied intrusions in the Rogaland complex^{7–9}. High-pressure melting studies, specifically on Harp Lake high-Al gabbros and Rogaland jotunitic rocks, have confirmed that these gabbroic/troctolitic to jotunitic suites fulfil the high- and low-pressure phase-equilibrium requirements as parental magmas for massif-type anorthosites³.

Conventional isotope tracers in anorthosite terranes have failed to resolve mantle versus lower crustal source regions, and generally yield ambiguous results. However, the basaltic affinity of the anorthosite suite, negatively correlated initial Sr isotope ratios and anorthite contents, and the bimodal nature of plutonism in many terranes have convinced most authors that the mantle is the most likely source^{2,10–12}. Plagioclase-saturated melts are not in

equilibrium with olivine at pressures exceeding ~9 kbar, so such magmas cannot be primary upper-mantle melts. If it is mantle-derived, the high-Al melt has to be the result of extensive fractionation of olivine and pyroxene from a more primitive magma at depth. On the basis of the low anorthite content and Mg number of massif-type anorthosite parental magmas, rare earth element modelling, and the somewhat erratic isotope signatures, some authors have argued for a lower crustal source^{13,14}. Lead isotope ratios have in some cases been interpreted as exhibiting crustal signatures¹ and initial Os isotopic compositions in sulphide ores from the Suwalki anorthosites in Poland are undeniably crustal¹⁵. Furthermore, a thermal barrier on the plagioclase-two pyroxene liquidus effectively prevents a mantle-derived basalt from evolving towards plagioclase-pyroxene saturation at depth³.

On a global scale, anorthosites and their associated primitive gabbroic to jotunitic suites display isotopic ratios with ϵ_{Nd} most often ranging from +5 to –7 and initial Sr isotope ratios from 0.703 to 0.706 (refs 4, 6, 11, 16). Positive to nearly chondritic ϵ_{Nd} -values have in many cases been taken as evidence for mantle derivation and the more negative values explained by assimilation of unradiogenic crustal material^{6,17}. However, even though the reduction in $^{147}\text{Sm}/^{144}\text{Nd}$ ratio in the lower crust is significant compared to unmetasomatized mantle, the lower crustal decrease in ϵ_{Nd} is slow. Lower crustal material derived from depleted mantle sources (characterized by positive ϵ_{Nd} values) may retain a positive ϵ_{Nd} for a considerable period of time. Given the variability in ϵ_{Nd} among various mantle reservoirs, this value on its own is not an effective discriminator between mantle and lower crustal reservoirs. Similarly, Rb-depletion of the lower crust effectively halts the Sr isotopic evolution and initial Sr isotope ratios will remain low. The Re–Os isotopic system is a more powerful tool to evaluate mantle versus lower crustal sources for mafic magmatism. Contrary to the Rb–Sr and Sm–Nd systems, mantle melting and crust formation fractionates Re from Os very efficiently, resulting in an increase of two orders of magnitude in $^{187}\text{Re}/^{188}\text{Os}$ in the lower crust relative to the upper mantle¹⁸. Even mafic lower crustal rocks with brief residence times will thus be characterized by very radiogenic Os isotopic signatures¹⁹.

We obtained whole-rock Re–Os isotopic analyses from a variety of rock types in the Rogaland anorthosite province. Our samples

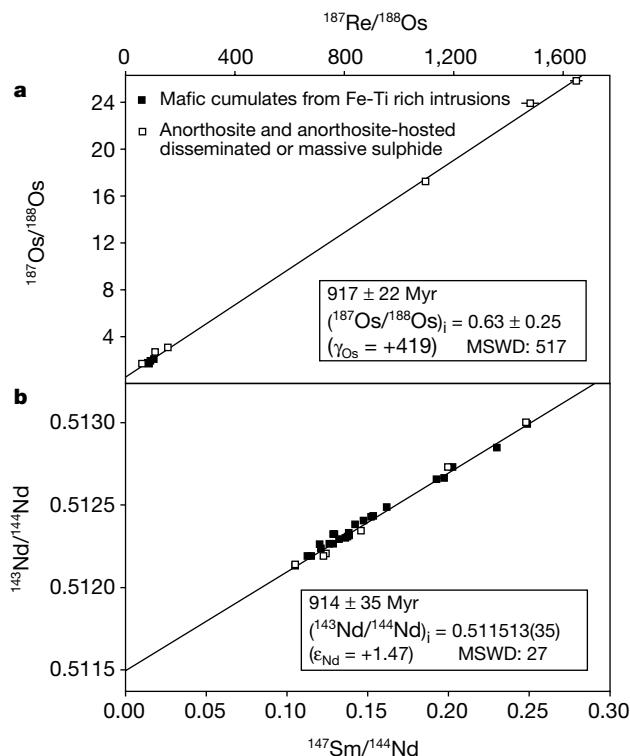


Figure 1 Isotopic data from Rogaland. **a**, Re–Os isotope data for ten samples of Rogaland massif-type anorthosite, anorthosite-hosted sulphide deposits, noritic and ilmenitic cumulates with a model 3 isochron. Error bars are shown when larger than symbol. The initial γ_{Os} value (that is, percent deviation from chondritic mantle at the time of formation) of +419 has been calculated for the U–Pb isotope age of 930 Myr (ref. 1). **b**, Sm–Nd isotope data for 30 samples from various intrusions in the Rogaland complex fitted with a model 3 isochron. The ϵ_{Nd} -value has been calculated for 930 Myr. For data and analytical details, see Supplementary Information. MSWD, mean square of weighted deviates.

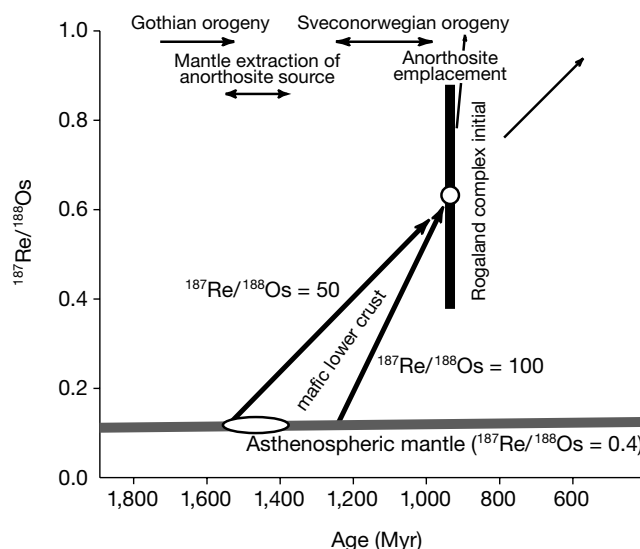


Figure 2 Modelling of crustal residence time for a lower crustal source derived from a mantle source with $\gamma_{\text{Os}} = 0$. Evolution of lower crust with $^{187}\text{Re}/^{188}\text{Os}$ ratios of 50 and 100 are shown along with the isotopic evolution of asthenospheric mantle. With lower crustal $^{187}\text{Re}/^{188}\text{Os}$ ratios between 50 and 70 the age of the inferred source for anorthosite magmatism is 1,550 to 1,400 Myr (open oval).

include unmineralized massif-type anorthosite and anorthosite-hosted sulphide deposits, high-temperature cumulates from the layered Bjerkreim–Sokndal Intrusion that crystallized after major replenishment events, and two ilmenite-rich cumulate dykes. Os concentrations vary from 0.005 p.p.b. in anorthosite to 10 p.p.b. in sulphide-bearing rocks (see Supplementary Information). On a Re–Os isotope evolution diagram (Fig. 1a) ten samples define an isochron of 917 ± 22 Myr, in agreement with high-precision U–Pb zircon and baddeleyite ages from the province between 920 and 930 Myr (ref. 1). Sm–Nd isotopic data from a similar variety of Rogaland rock types (exclusive of granitoids) corroborate the Re–Os isochron age (Fig. 1b). Thirty analyses fit an isochron of 914 ± 35 Myr with an initial ϵ_{Nd} of +1.47.

It is apparent from Fig. 1 that the source characteristics for the massif-type anorthosites, the internal sulphides and the associated noritic and ilmenite-rich rocks are alike. Each intrusion is characterized by similar initial Sr isotope ratios, and mineral chemistry and phase assemblages are correlated throughout the province. The observations indicate that all intrusions were derived from discrete melt batches of similar composition. Because our samples were collected from several intrusions up to 50 km apart, source heterogeneity is the most probable cause of the high mean square of weighted deviates recorded by both isochrons. The isochronous relationships and concordant isochron ages documented by the two very different geochronometers eliminates the possibility of variable amounts of crustal contamination of mantle-derived melts. The isochron initial $^{187}\text{Os}/^{188}\text{Os}$ ratio of 0.63 ± 0.25 corresponds to a γ_{Os} value (that is, percent deviation from chondritic mantle at the time

of formation) of +419 (at 930 Myr). As the upper mantle is characterized by γ_{Os} near 0 (± 15), the high initial γ_{Os} demonstrates the involvement of a major crustal component in the genesis of the anorthosites and related rocks.

Estimates of the Re–Os composition of the crust are sparse. An average upper-crustal $^{187}\text{Re}/^{188}\text{Os}$ ratio of 42–44 is generally accepted²⁰. The highest Re/Os ratios are normally encountered in mafic systems whereas felsic intrusives such as granites and tonalites are characterized by lower $^{187}\text{Re}/^{188}\text{Os}$ ratios and lower Re and Os concentrations than the upper crustal average^{19,21}. Lower crustal xenoliths and granulitic representatives of lower crust have variable $^{187}\text{Re}/^{188}\text{Os}$ ratios generally between 10 and 100 (refs 19, 22–24), and locally higher.

If the source of the massif-type anorthosites and their related primitive intrusives is located in the lower crust^{13,14} we may thus adopt a $^{187}\text{Re}/^{188}\text{Os}$ ratio of 50 as a likely composition of the source region. A mafic lower crust with such a $^{187}\text{Re}/^{188}\text{Os}$ ratio requires 610 Myr to evolve from $\gamma_{\text{Os}} = 0$ to +419 (Fig. 2), implying a crustal source 1,540 Myr old. This is significant as the earliest and major crust-forming event in southwest Norway took place during the Gothian orogeny at about 1,570 Myr ago¹¹. There is no evidence from regional surveys and isotopic mantle-extraction ages for the presence of pre-Gothian crust in the Rogaland–Vest Agder tectonic segment hosting the anorthosite province; the oldest age reported is 1,450 Myr (ref. 25). Increasing the assumed lower crustal $^{187}\text{Re}/^{188}\text{Os}$ to 100 results in a modelled crustal residence time of 306 Myr, corresponding to an age of 1,236 Myr.

Initial ϵ_{Nd} values for Rogaland are restricted to +0.22–2.04, ruling out both parental magmas and major crustal contaminants extracted from primitive mantle sources (Fig. 3). Depleted mantle will be characterized by an elevated ϵ_{Nd} of around +4.3 at 1,540 Myr and +6.3 at 930 Myr (ref. 26). The $^{147}\text{Sm}/^{143}\text{Nd}$ ratio of the lower crust is variable and poorly constrained but the values obtained in the anorthosite province have to represent a minimum for the underlying mafic crust. This means that the $^{147}\text{Sm}/^{144}\text{Nd}$ ratio of the lower crust in the province has to be higher than 0.13. Assuming $^{147}\text{Sm}/^{144}\text{Nd}$ of 0.15 gives a model age of 1,440 Myr for the lower crustal source.

An explanation of the Rogaland Os and Nd initial ratios by crustal contamination of mantle-derived magma is implausible, as may be demonstrated by mixing models, adopting the most favourable scenario possible for the Rogaland province (Fig. 3). Even in this extreme case, a depleted mantle-derived melt would have to assimilate at least 30% crustal material to account for the observed ratios, and the concentration of Os in the contaminant needs to be approximately twice the concentration in the mantle melt. The latter is improbable as, in general, the Os concentration in the lower crust is similar to or lower than in a mantle-derived melt, depending on the degree of partial melting²⁷. The required Os-content (≥ 0.2 p.p.b.) implies a very refractory mafic contaminant, which is not consistent with a large amount of assimilation. Also, because Os abundance and Re/Os ratios are negatively correlated in mantle-derived melts and their bulk crustal derivatives, the coupled requirement for high Os and high Re/Os necessitates that the contaminant involved is unusual. The absence in the region of older, more radiogenic crust therefore effectively rules out the hypothesis of lower crustal contamination of mantle-derived magmas to account for the initial ratios.

The Sm–Nd and Re–Os modelling therefore indicates a common origin of the anorthosites, their satellite intrusions and a magma source located in Mesoproterozoic mafic lower crustal rocks. Initial lead isotopic compositions of plagioclase and inherited zircon components from the Rogaland complex, also indicate a crustal source of middle to late Mesoproterozoic age for the anorthosite province¹.

In Rogaland, the layered Bjerkreim–Sokndal Intrusion demonstrably crystallized from less-differentiated equivalents of the

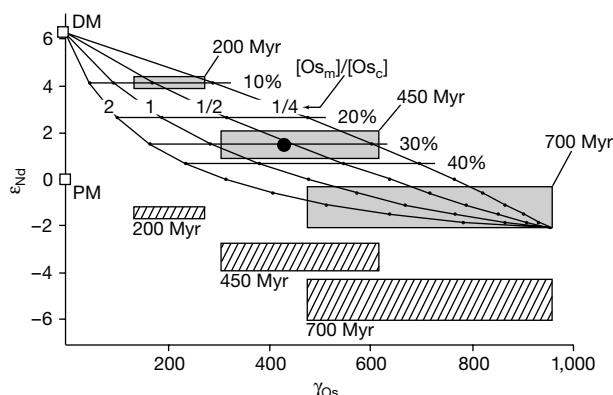


Figure 3 Modelling of a primary lower crustal source versus lower crustal contamination of a primary mantle-derived magma in $\gamma_{\text{Os}}-\epsilon_{\text{Nd}}$ space. Grey and hatched squares show the composition of crustal rocks 200, 450 and 700 Myr old at 930 Myr ago, derived from either depleted (grey) or primitive (CHUR-type) mantle. DM, depleted mantle; PM, primitive mantle. The compositional span is calculated using $^{187}\text{Re}/^{188}\text{Os} = 50-100$ and $^{147}\text{Sm}/^{144}\text{Nd} = 0.13-0.15$. The black dot is the initial Nd and Os isotopic composition of the Rogaland complex, similar to that expected for ~450-Myr-old crust derived from depleted mantle. Models invoking primitive-mantle sources for melts or contaminants are not consistent with the Rogaland data. Mixing hyperbolae are shown for the most extreme scenario possible in the Rogaland province, with a melt derived from depleted mantle assimilating rocks of the highest possible age (700 Myr old), the highest $^{187}\text{Re}/^{188}\text{Os}$ (100) and the lowest $^{147}\text{Sm}/^{144}\text{Nd}$ (0.13) within the possible range. Mixing curves are modelled for different melt-contaminant Os-concentration ratios ($[\text{Os}_{\text{melt}}]/[\text{Os}_{\text{contaminant}}]$) and the degree of assimilation is shown in percent. The model indicates the possibility of the initial Nd and Os isotopic composition of the Rogaland complex originating by 30% assimilation of old, compositionally extreme crust by a depleted mantle-derived melt, provided that the crustal contaminant contains twice as much Os as the primary melt, which is an unlikely eventuality (see text). In addition, 30% contamination is likely to erase all isochronous relationships. A constant $[\text{Nd}_{\text{melt}}]/[\text{Nd}_{\text{contaminant}}] = 3$ is used in the models. Changing this ratio will result in either higher amounts of assimilation or higher Os-concentrations in the contaminant.

jotunitic chilled margins^{8,9}. Chilled margins of similar composition are also found around the Hydra anorthosite body⁷. Batches of primitive jotunitic magma derived by partial fusion of a mafic lower crust of Gothian to post-Gothian age (1,550–1,400 Myr ago) are therefore the best candidates for parental melts, not only for the massif-type anorthosites but also for the associated noritic and ilmenite-rich intrusions in the Rogaland anorthosite province. The variations in parental magma composition inferred for other anorthosite provinces probably depend on the exact nature of the underlying crust.

The Rogaland anorthosites were emplaced post-orogenically ~60 Myr after the latest recorded Sveconorwegian deformation, coinciding with post-orogenic granitoid magmatism elsewhere in south Norway and western Sweden. Seismic and gravimetric profiles along the south coast of Norway have shown that Sveconorwegian Moho-ramp tectonics cause the intrusion of crustal tongues into the contemporary mantle²⁸. Melting of these crustal tongues was recently proposed to account for the anorthosite formation²⁹ and currently provides us with the most plausible setting for lower crustal melting and genesis of the parental magmas for the Rogaland massif-type anorthosites. □

Received 4 January; accepted 25 April 2000.

- Schärer, U., Wilms, E. & Duchesne, J. C. The short duration and anorogenic character of anorthosite magmatism: U-Pb dating of the Rogaland complex, Norway. *Earth Planet. Sci. Lett.* **139**, 335–350 (1996).
- Emslie, R. F. In *The Deep Proterozoic Crust in the North Atlantic Provinces* (eds Tobin, A. C. & Touret, J. L. R.) 39–60 (Reidel, Dordrecht, 1985).
- Longhi, J., Vander Auwera, J., Fram, M. S. & Duchesne, J. C. Some phase equilibrium constraints on the origin of Proterozoic (massif) anorthosites and related rocks. *J. Petrol.* **40**, 339–362 (1999).
- Scotese, J. S. & Frost, C. D. A strontium and neodymium isotopic investigation of the Laramie anorthosites, Wyoming, USA: Implications for magma chamber processes and the evolution of magma conduits in Proterozoic anorthosites. *Geochim. Cosmochim. Acta* **60**, 95–107 (1996).
- Longhi, J., Fram, M. S., Vander Auwera, J. & Monteth, J. N. Pressure effects, kinetics, and rheology of anorthositic and related magmas. *Am. Mineral.* **78**, 1016–1030 (1993).
- Mitchell, J. M., Scotese, J. S. & Frost, C. D. High-Al gabbros in the Laramie Anorthosite Complex, Wyoming: implications for the composition of melts parental to Proterozoic anorthosite. *Contrib. Mineral. Petrol.* **119**, 166–180 (1995).
- Demaiffe, D. & Hertogen, J. Rare earth element geochemistry and strontium isotopic composition of a massif-type anorthositic-charnockitic body: the Hydra Massif (Rogaland, SW Norway). *Geochim. Cosmochim. Acta* **45**, 1545–1561 (1981).
- Robins, B., Tumyr, O., Tysse, L. M. & Garmann, L. B. The Bjerkreim-Sokndal Layered Intrusion, Rogaland, SW Norway: Evidence from marginal rocks for a jotunitic parent magma. *Lithos* **39**, 121–133 (1997).
- Duchesne, J. C. & Hertogen, J. Le magma parental du lopolithe de Bjerkreim-Sokndal (Norvège méridionale). *C. R. Acad. Sci.* **306**, 45–48 (1988).
- Morse, S. A. A partisan review of Proterozoic anorthosites. *Am. Min.* **67**, 1087–1100 (1982).
- Menage, J. F. The petrogenesis of massif anorthosites: a Nd and Sr isotopic investigation of the Proterozoic of Rogaland/Vest-Agder, SW Norway. *Contrib. Mineral. Petrol.* **98**, 363–373 (1988).
- Ashwal, L. D. & Seifert, K. E. Rare-earth-element geochemistry of anorthosite and related rocks from the Adirondacks, New York, and other massif-type complexes: Summary. *Geol. Soc. Am. Bull.* **91**, 105–107 (1980).
- Simmons, E. C. & Hanson, G. N. Geochemistry and origin of massif-type anorthosites. *Contrib. Mineral. Petrol.* **66**, 119–135 (1978).
- Taylor, S. R., Campbell, I. H., McCulloch, M. T. & McLennan, S. M. A lower crustal origin for massif-type anorthosites. *Nature* **311**, 372–374 (1984).
- Stein, H. J., Morgan, J. W., Markey, R. J. & Wisniewska, J. A Re-Os study of the Suwalki Anorthosite Massif, Northeast Poland. *Geophys. J.* **4**, 111–114 (1998).
- Demaiffe, D., Weis, D., Michot, J. & Duchesne, J. C. Isotopic constraints on the genesis of the Rogaland anorthositic suite (Southwest Norway). *Chem. Geol.* **57**, 167–179 (1986).
- Emslie, R. F., Hamilton, M. A. & Thériault, R. J. Petrogenesis of a Mid-Proterozoic anorthosite-mangerite-charnockite-granite (AMCG) complex: Isotopic and chemical evidence from the Nain Plutonic Suite. *J. Geol.* **102**, 539–558 (1994).
- Walker, R. J., Carlson, R. W., Shirey, S. B. & Boyd, F. R. Os, Sr, Nd and Pb isotope systematics of southern African peridotite xenoliths: Implications for the chemical evolution of subcontinental lithospheric mantle. *Geochim. Cosmochim. Acta* **53**, 1583–1595 (1989).
- Asmerom, Y. & Walker, R. J. Pb and Os isotopic constraints on the composition and rheology of the lower crust. *Geology* **26**, 359–362 (1998).
- Esser, B. K. & Turekian, K. K. The osmium isotopic composition of the continental crust. *Geochim. Cosmochim. Acta* **57**, 3093–3104.
- Johnson, C. M., Shirey, S. B. & Barovich, K. M. New approaches to crustal evolution studies and the origin of granitic rocks: what can the Lu-Hf and Re-Os isotope systems tell us? *Trans. R. Soc. Edinb.* **87**, 339–352.
- Esperança, S., Carlson, R. W., Shirey, S. B. & Smith, D. Dating crust-mantle separation: Re-Os isotopic study of mafic xenoliths from central Arizona. *Geology* **25**, 651–654 (1997).
- Saal, A. E., Rudnick, R. L., Ravizza, G. E. & Hart, S. R. Re-Os isotopic evidence for the composition, formation and age of the lower continental crust. *Nature* **393**, 58–61 (1998).
- Frick, L. R., Lambert, D. D. & Cartwright, I. Re-Os dating of metamorphism in the Lewisian Complex, NW Scotland. *Goldschmidt Symp. (Heidelberg)* *J. Conf. Abs.* **1**, 85 (1996).

- Verstevee, A. J. Isotope geochronology in the high-grade metamorphic Precambrian of Southwestern Norway. *Geol. Surv. Norway Bull.* **318**, 1–50 (1975).
- Michard, A., Gurriet, P., Soudant, M. & Albareda, F. Nd isotopes in French Phanerozoic shales: external vs. internal aspects of crustal evolution. *Geochim. Cosmochim. Acta* **49**, 601–610 (1985).
- Lambert, D. D., Foster, J. G., Frick, L. R., Ripley, E. M. & Zientek, M. L. Geodynamics of magmatic Cu-Ni-PGE sulfide deposits: New insights from the Re-Os isotope system. *Econ. Geol.* **93**, 121–135 (1998).
- Andersson, M., Lie, J. E. & Husebye, E. S. Tectonic setting of post-orogenic granites within SW Fennoscandia based on deep seismic and gravity data. *Terra Nova* **8**, 558–566 (1996).
- Duchesne, J. C., Liégeois, J. P., Vander Auwera, J. & Longhi, J. The crustal tongue melting model and the origin of massive anorthosites. *Terra Nova* **11**, 100–105 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

Financial support for Re-Os isotopic studies at Monash University derive from the Monash University Research Fund, the Australian Crustal Research Centre, and the Australian Research Council. The research reported here was supported by the Norwegian Research Council.

Correspondence and requests for materials should be addressed to H.S. (e-mail: henrik.schiellerup@geo.ntnu.no).

Unrelated helpers in a social insect

David C. Queller*, Francesca Zacchi*, Rita Cervo†, Stefano Turillazzi†, Michael T. Henshaw*, Lorenzo A. Santorelli† & Joan E. Strassmann*

* Department of Ecology and Evolution, Rice University, PO Box 1892, Houston, Texas 77251-1892, USA

† Dipartimento di Biologia Animale e Genetica, Università di Firenze, Via Romana 17, 50125 Florence, Italy

High-resolution genetic markers have revolutionized our understanding of vertebrate mating systems¹, but have so far yielded few comparable surprises about kinship in social insects. Here we use microsatellite markers to reveal an unexpected and unique social system in what is probably the best-studied social wasp, *Polistes dominulus*. Social insect colonies are nearly always composed of close relatives^{2,3}; therefore, non-reproductive helping behaviour can be favoured by kin selection, because the helpers aid reproductives who share their genes⁴. In *P. dominulus*, however, 35% of foundress nestmates are unrelated and gain no such advantage. The *P. dominulus* system is unlike all other cases of unrelated social insects, because one individual has nearly

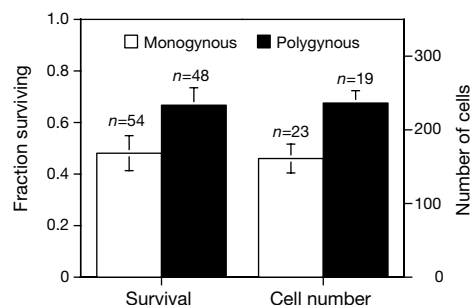


Figure 1 Success of colonies begun by one foundress and multiple foundresses (monogynous and polygynous, respectively) measured for 1995 colonies through mid-August, when reproductives are being produced. Polygynous colonies are more successful both for survival ($P < 0.01$, test for equality of percentages) and cell number ($P < 0.01$, t -test). Relative numbers of monogynous and polygynous colonies in the survival study are representative of its source population. The cell number data excludes failed nests and nests rebuilt after predation. Bars show standard error (s.e.).

complete reproductive dominance over subordinates who could have chosen other reproductive options. The only significant advantage that subordinates obtain is a chance at later reproduction, particularly if the queen dies. Thus, *P. dominulus* societies are functionally unlike other social insects, but similar to certain vertebrate societies^{5,6}, in which the unrelated helpers gain through inheritance of a territory or a mate.

Polistes dominulus (formerly known as *Polistes gallicus*) has a typical *Polistes* colony cycle⁷. Colonies are begun in the spring by one or more mated, overwintered foundresses. When there is more than one, a hierarchy emerges⁸ (this is the first invertebrate species in which dominance hierarchies were described). One foundress becomes the dominant queen and lays most of the eggs, while the subordinates take on most of the riskier foraging work, substantially increasing colony success⁹ (Fig. 1). The first brood emerge as adults in June and are mostly female workers (the few males produced may mate with workers fortunate enough to assume queenship when foundresses all die). The workers help to rear additional brood until the season ends with the rearing of males and gynes¹, the non-worker females that will be the next year's foundresses.

We collected four sets of multiple-foundress colonies from open-ended plastic tubes protecting saplings in fields near Cavriglia in central Italy: two 'early foundress' sets in April shortly after nest founding; and two 'late foundress' sets in May or June, shortly before worker emergence (Fig. 2). We genotyped six or seven microsatellite loci for all foundresses, and for samples of brood in the two 1996 collections (eggs for 'early'; eggs and pupae for 'late').

Average relatedness among nestmate foundresses was low in all four collections (Fig. 2). The distribution of relatedness estimates shows a large peak at the haplodiploid full-sister value of three-quarters, and another substantial peak near zero (Fig. 3). The expected distributions of relatedness estimates overlap broadly for non-relatives and cousins (Fig. 3), but a likelihood analysis shows that many, probably most, of these are truly non-relatives. The maximum likelihood of obtaining the observed distribution is achieved when 35% are drawn from the non-relative distribution, 9% from the cousin distribution and 56% from the full-sister

distribution. This hypothesis was more than twenty times more likely than any hypothesis with fewer than 20% non-relatives. Thus, the previously noted tolerance of non-nestmate foundresses forced together in the lab^{10,11} is not simply a laboratory artefact but a real feature of the life history.

P. dominulus does not match either of the two patterns common to other unrelated social insects. Unrelated ant foundresses¹² and communal insects¹³ do not accept a deeply subordinate role. Instead, reproduction and labour are divided relatively equally (although ant foundresses usually fight to the death after worker emergence). One ant, *Acromyrmex versicolor*, shows division of labour among unrelated foundresses, but foragers apparently also reproduce¹⁴. In contrast, *P. dominulus* subordinates hardly reproduce at all in the presence of the queen. The queen was the source of most genotyped eggs assigned to living females: 93.9% among early foundresses and 99.6% in late foundresses (1996 collections; six colonies in which the largest egg producer was dead were omitted because one of the apparent subordinates may have been reproducing as the new queen). Very few brood were attributed to the average subordinate collected (early, 2.6%; late, 0.2%; a few others were attributed to dead wasps).

The few subordinates that reproduced (four in April, one in June) were not less related to their queen than non-reproducers ($r = 0.58$ versus 0.42, two-tailed $t = 0.79$, $P = 0.43$), contrary to a reproductive skew theory that predicts that the controlling dominant must cede more reproduction to unrelated helpers to secure their help^{7,15,16}. Skew theory would be irrelevant to purely non-reproductive brood, but workers do reproduce if foundresses die¹⁷ (whether they also produce males in the presence of foundresses is unknown).

The other known pattern is that unrelated individuals may help when there are no other real options. Examples include workers of uniclonal ants¹⁸ and *Polistes* females whose nests are forcibly usurped too late in the season to begin anew^{7,19}. In contrast, *P. dominulus* subordinates working for unrelated queens do have other options. Our early foundress collections were from a time when colonies were still being initiated, so leaving a usurped colony is still an option. Solitary nesting is a common and viable strategy⁹ (Fig. 1) and there were many vacant tubes available for nesting. Usurpation is also ruled out as a cause of unrelated foundress associations by censuses of the individually marked 1996 late foundresses (nine censuses over five weeks, conducted in early morning, when all wasps are present). Of 49 subordinates collected, only 6 were observed on the nest before their queen, and these were not less related to their queen than were other subordinates ($r = 0.36$ versus 0.32). A second option, joining a full-sister queen on another nest, is clearly also available for many unrelated subordinates²⁰, but

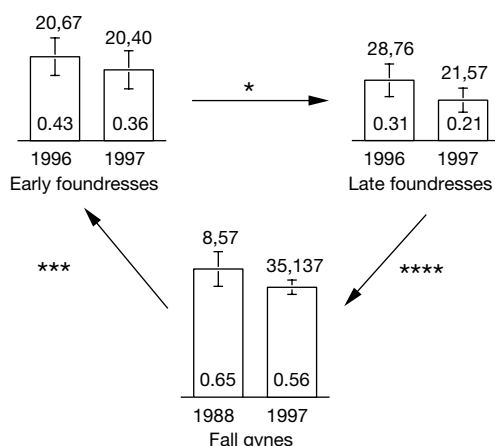


Figure 2 Relatedness ($\pm$ s.e.) in three categories of adult female nestmates (the 1988 autumn gyne value is from a previously published allozyme study³⁰). Between categories, each asterisk represents one significant difference from among the four possible comparisons ($P < 0.05$, one-tailed t -tests). There are no significant differences between collections within categories. Numbers above bars are colonies and individuals genotyped. As all collected foundresses (but not autumn gynes) were genotyped, these numbers yield the average association size that was collected, excluding single foundress colonies. Because association size was not significantly linked to relatedness in 1996, we chose 1997 foundress associations that were slightly smaller than average to facilitate behavioural studies.

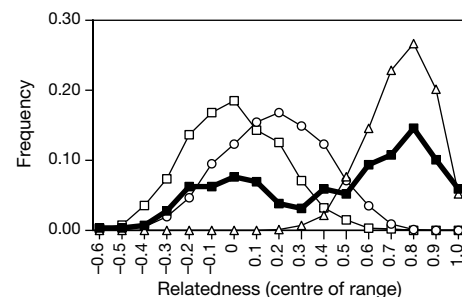


Figure 3 Observed relatedness distribution of *P. dominulus* foundresses (filled squares) and expected distributions for several relationships, grouped into intervals of width 0.1. The filled squares show the observed distribution for all relatedness estimates of foundress nestmate pairs, from the four foundress collections of Fig. 2. The other distributions, used in the likelihood analysis, show the distributions of relatedness estimates for simulated non-relatives (open squares, true $r = 0$), cousins (open circles, true $r = 3/16$) and full sisters (open triangles; true $r = 3/4$).

it is difficult to quantify because of the uncertainty of individual relatedness estimates (Fig. 3) and because we did not collect all nearby nests.

Why do unrelated subordinates help if they obtain little direct or indirect reproduction? Such help could be maladaptive, but we found no evidence for errors in homing or recognition. If homing errors were common, we would expect to see considerable movement among nests, but in the 9 censuses of the 1996 late foundresses, only 3 of the 123 marked foundresses were ever seen on more than one nest (although the nest tubes were similar, they were on trees of different species and shapes, and nests were built at different positions within their tubes). Nestmate recognition abilities are intact in these populations; wasps artificially transferred between nests after workers have emerged are vigorously attacked. We cannot exclude the possibility that recognition is somehow lost only by overwintering foundresses, but it would raise the mystery of why other *Polistes* do not share this problem.

Finally, unrelated foundresses might treat each other as related if they emerge from the same autumn nest, because that is where they learn recognition cues²¹. A likelihood analysis exactly parallel to that above suggests that 12% of the autumn 1997 gynes are indeed unrelated. However, this cannot account for the significantly lower relatedness among spring foundresses (Fig. 2; the unweighted relatedness differences between autumn gynes and spring foundresses could be inflated if colonies with many gynes had lower relatedness, but they did not; Spearman's test $r = 0.066$, $P = 0.77$, $n = 23$ colonies from 1997). A joint likelihood analysis (of all the populations analysed by microsatellites) shows that it is over 300 times more likely that the spring and autumn collections came from separate distributions (with 35% and 12% unrelated pairs) than from the most probable single distribution (24% unrelated).

The remaining possible advantage is delayed reproduction, perhaps through inheritance of the nest and worker force. Of the 28 colonies collected in the late foundress stage in 1996, 11 showed a total of at least 13 queen turnovers. The new queens were not generally usurpers: 10 out of 13 had been present at the very first nest census, a month or more before collection. Thus 11 of the 110 censused subordinates had become dominant at the time of collection, compared with 17 of the 28 original queens. The average subordinate is therefore 16.5% as likely to be queen just before worker emergence as is an original queen. This is an order of magnitude higher than the subordinates' relative egg-laying success in the presence of the queen (see above; the fitness difference is even larger, as late reproduction is more important than the early reproduction that produces mostly workers). After worker emergence further colony inheritance by subordinates may occur, but even without it, the primary benefit of unrelated foundresses is delayed reproduction. Delayed reproduction may also account for the apparent failure of the skew theory noted above. When delayed reproduction is sufficiently valuable, the queen may not need to cede any reproduction to either related or unrelated subordinates^{22,23}.

Relatedness did not clearly affect either colony growth (Fig. 4c) or behaviour in laboratory colonies (Fig. 4a and b), although the tendency for close relatives to spend more time off the nest (Fig. 4a) may be worth further study in the field, where such time is spent in hazardous foraging. Although one might expect non-relatives to work less to increase their chance of inheritance⁵, it may be that working is the price they are forced to pay to remain.

No other social insect is known to accept such profound submission to non-relatives when other options are available. While wider application of highly variable markers could reveal that this pattern is more common, the closest known parallel is to a few vertebrate social systems, in which both related and unrelated subordinates help, the latter gaining only by the chance to inherit the territory or a mate^{5,6}. □

Methods

Microsatellite genotyping

We genotyped 6–7 microsatellite loci, previously developed for other species (Paco3434AAT²⁴, Pan80AAT²⁵, Pbe102TAG, Pbe128TAG, Pbe269AAG, Pbe430AAG, Pbe440AAG²⁶) using standard methods²⁷. Two people independently scored gels and entered data, and discrepancies were resolved by re-examination of the gel and, if necessary, re-running the sample. Expected heterozygosities ranged from 0.63 to 0.89.

Relatedness and brood assignments

We estimated relatedness²⁸ using the computer program Relatedness 5.05 (<http://gsoft.smu.edu/GSoft.html>). We used jack-knife calculations over colonies to obtain standard errors and *t*-tests.

To estimate fractions of full sisters, cousins and non-relatives, we simulated 2,000 pairs of each, drawn from a population with the observed microsatellite allele frequencies (using the Macintosh program Kinship 1.3 (<http://gsoft.smu.edu/GSoft.html>)). We then calculated the likelihood of drawing our observed relatedness distribution from any mixture of the three simulated distributions (the simulations omitted the seventh locus used in only two studies; this broadens the distributions, making the likelihood analysis more conservative).

We assigned brood to foundresses by first grouping all brood into full-sister groups, which at every locus shared one allele (from the haploid father) and collectively had no more than two other alleles (from the diploid mother). These full-sister groups did not aggregate into larger groups attributable to the same mother, indicating effective single mating. Single mating was confirmed by genotyping the sperm in foundress spermathecae (although many cases failed because of either maternal contamination or amplification failure²⁰). We then matched the parental genotypes inferred from the sister groups to observed maternal genotypes (and where available, sperm genotypes). This procedure is justified by the fact that matching was usually unique, and the largest full-sister group matched the foundress with the best developed ovaries, unless it matched no foundress. In the latter case, we inferred that the queen had disappeared, and reconstructed her genotype at as many loci as possible.

We detected queen turnovers in ten colonies where the foundress who dominated old brood (pupae) production was different from the one dominating young brood (egg) production (in nine of these cases the difference was significant; Fisher's exact test). Three additional queen disappearances were inferred when the foundress dominating egg production disappeared before collection (total eggs laid by collected foundresses: 0/6, 0/9, 1/6). On these three nests, all collected foundresses had been censused on the first census

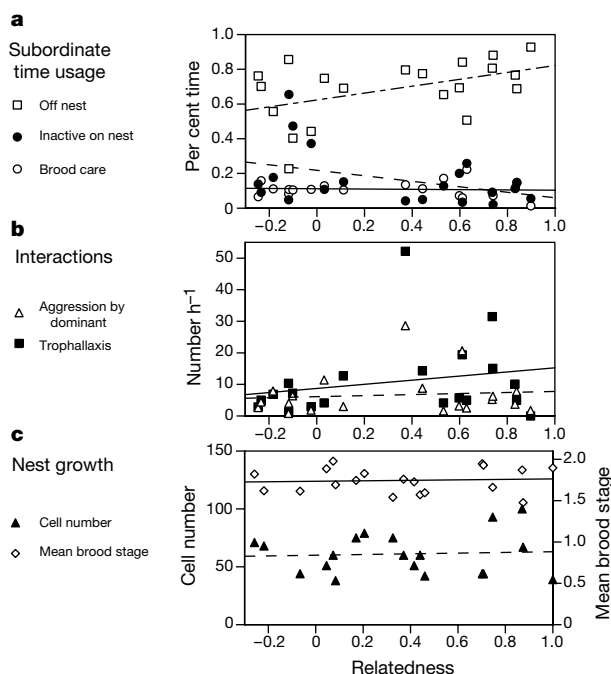


Figure 4 Regressions of behaviour and colony growth on foundress relatedness. **a, b**, Data taken from 20 two-foundress colonies (1997 early foundress collection); **c**, data taken from colonies of the 1996 late foundress collection (here mean relatedness is used). Mean brood stage is calculated by assigning eggs a value of 1, larvae 2 and pupae 3. Two regressions are significant individually (subordinate's time off the nest, $P = 0.02$; and subordinate's time inactive on the nest, $P = 0.04$), but not when corrected for multiple comparisons (five behavioural tests). All tests are one-tailed, with less-related subordinates predicted to work less and be subject to more aggression.

date, so the successor, although unknown, had to be an established foundress rather than a recent usurper.

Lab observations

We set up 20 two-foundress colonies, including the nest with brood, in individual glass and wire-screen enclosures (15 cm)³. We kept them under incandescent lighting 12 h d⁻¹, and provided water, honey and *Tenebrio molitor* larvae *ad lib*. Acclimatization for 2–7 d preceded videotaping for 11.8 ± 0.2 h (mean ± s.d.). All visible instances of the behaviours in Fig. 4 were scored. Interactions between dominant and subordinate were calculated per hour when both were on the nest. Aggression includes climbing on top, chewing and lunging.

Received 28 January; accepted 15 March 2000.

1. Hughes, C. R. Integrating molecular techniques with field methods in studies of social behavior: a revolution results. *Ecology* **79**, 383–399 (1998).
2. Crozier, R. H. & Pamilo, P. *Evolution of Social Insect Colonies: Sex Allocation and Kin Selection* (Oxford Univ. Press, Oxford, 1996).
3. Queller, D. C. & Strassmann, J. E. Kin selection and social insects. *Bioscience* **48**, 165–174 (1998).
4. Hamilton, W. D. The genetical evolution of social behaviour. I, II. *J. Theor. Biol.* **7**, 1–52 (1964).
5. Reyer, H.-U. Investment and relatedness: a cost/benefit analysis of breeding and helping in the pied kingfisher (*Ceryle rudis*). *Anim. Behav.* **32**, 1163–1178 (1984).
6. Creel, S. R. & Waser, P. M. Inclusive fitness and reproductive strategies in dwarf mongooses. *Behav. Ecol.* **5**, 339–348 (1994).
7. Reeve, H. K. in *The Social Biology of Wasps* (eds Ross, K. G. & Matthews R. W.) 99–148 (Cornell Univ. Press, Ithaca, 1991).
8. Pardi, L. Dominance order in *Polistes* wasps. *Physiol. Zool.* **21**, 1–13 (1948).
9. Turillazzi, S., Marino Piccoli, M. T., Hervatin, L. & Pardi, L. Reproductive capacity of single foundress and associated foundress females of *Polistes gallicus* (L.) (Hymenoptera Vespidae). *Monitore Zool. Ital. (N. S.)* **16**, 75–88 (1982).
10. Pratte, M. Relations antérieures et association de fondation chez *Polistes gallicus* L. *Insectes Sociaux* **29**, 352–357 (1982).
11. Röseler, P. F. in *Experimental Behavioral Ecology and Sociobiology* (eds Hölldobler, B. & Lindauer, M.) 259–272 (Sinauer, Sunderland, 1985).
12. Bernasconi, G. & Strassmann, J. E. Cooperation among unrelated individuals: the ant foundress case. *Trends Ecol. Evol.* **14**, 477–482 (1999).
13. Kukuk, P. F. & Sage, G. K. Reproductivity and relatedness in a communal halictine bee, *Lasiglossus (Chilalictus) hemichalceus*. *Insectes Sociaux* **41**, 443–455 (1994).
14. Rissing, S. W., Pollock, G. B., Higgins, M. R., Hagen, R. H. & Smith, D. R. Foraging specialization without relatedness or dominance among co-founding ant queens. *Nature* **337**, 420–422 (1989).
15. Vehrencamp, S. L. Optimal degree of skew in cooperative societies. *Am. Zool.* **23**, 327–335 (1983).
16. Reeve, H. K. & Ratnieks, F. L. W. in *Queen Number and Sociality in Insects* (ed Keller, L.) 45–85 (Oxford Univ. Press, Oxford, 1993).
17. Pardi, L. Ricerche sui polistini 11. Sulla durata permanenze della femmine nel nido e sul accrescimento della società in *Polistes gallicus* (L.). *Atti Soc. Toscana Sci. Nat.* **55**, 3–15 (1948).
18. Bourke, A. F. G. & Franks, N. R. *Social Evolution in Ants* (Princeton Univ. Press, Princeton, USA, 1995).
19. Lorenzi, M. C. & Cervo, R. Usurpations and late associations in the solitary founding social wasp, *Polistes biglumis bimaculatus* (Hymenoptera: Vespidae). *J. Insect Behav.* **8**, 443–451 (1995).
20. Zacchi, F. Genetic structure of *Polistes dominulus* foundress associations. PhD thesis, Rice University, Houston (1998).
21. Gamboa, G. J., Reeve, H. K. & Pfennig, D. W. The evolution and ontogeny of nestmate recognition in social wasps. *Annu. Rev. Entomol.* **31**, 431–454 (1986).
22. Kokko, H. & Johnstone, R. A. Social queuing in animal societies: a dynamic model of reproductive skew. *Proc. R. Soc. Lond. B* **266**, 571–578 (1999).
23. Ragsdale, J. E. Reproductive skew theory extended: the effect of resource inheritance on social organization. *Evol. Ecol. Res.* **1**, 859–874 (1999).
24. Strassmann, J. E., Solis, C. R., Barefield, K. & Queller, D. C. Trinucleotide microsatellite loci in a swarm-founding neotropical wasp, *Parachartergus colobopterus* and their usefulness in other social wasps. *Mol. Ecol.* **5**, 459–461 (1996).
25. Strassmann, J. E. *et al.* Trinucleotide microsatellite loci and increased heterozygosity in cross-species applications in the social wasp, *Polistes*. *Biochem. Genet.* **35**, 273–279 (1997).
26. Strassmann, J. E., Barefield, K., Solis, C. R., Hughes, C. R. & Queller, D. C. Trinucleotide microsatellite loci for a social wasp, *Polistes*. *Molec. Ecol.* **6**, 97–100 (1997).
27. Strassmann, J. E., Solis, C. R., Peters, J. M. & Queller, D. C. in *Molecular Zoology: Advances, Strategies and Protocols* (eds Ferraris, J. D. & Palumbi, S. R.) 163–180, 528–549 (Wiley-Liss, New York, NY, 1996).
28. Queller, D. C. & Goodnight, K. F. Estimating relatedness using genetic markers. *Evolution* **43**, 258–275 (1989).
29. Goodnight, K. F. & Queller, D. C. Computer software for performing likelihood tests of pedigree relationship using genetic markers. *Mol. Ecol.* **8**, 1231–1234 (1999).
30. Strassmann, J. E. *et al.* Genetic relatedness in primitively eusocial wasps. *Nature* **342**, 268–270 (1989).

Acknowledgements

We thank G. Bernasconi and L. Keller for helpful comments on the manuscript and ENEL SpA PDT Centro for the permission to work in the re-forestation area of the 'Miniera di Santa Barbara'. This work was supported by the NSF (US) and the Italian MURST. We dedicate this paper to the memory of Bill Hamilton, who taught us how to think about social insect evolution.

Correspondence and requests for materials should be addressed to D.C.Q. (e-mail: queller@rice.edu).

Female feral fowl eject sperm of subordinate males

T. Pizzari & T. R. Birkhead

Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

Paternity is often determined by competition between the ejaculates of different males¹. Males can also use particular behaviours or structures to manipulate how females use sperm^{2–5}. However, the ability of females to bias sperm utilization in favour of preferred males independently of male manipulation has not been demonstrated⁶. Females are predicted to respond differentially to the sperm of different males when the reproductive interests of the sexes differ and when females are coerced into copulating^{4,6}. Here we show that in female feral fowl most copulations are coerced, and that females consistently bias sperm retention in favour of the preferred male phenotype. Females prefer to copulate with dominant males, but when sexually coerced by subordinate males, they manipulate the behaviour of dominant males to reduce the likelihood of insemination. If this fails, females differentially eject ejaculates according to male status in the absence of any male manipulation and preferentially retain the sperm of dominant males.

Females can influence paternity before insemination by selecting their copulation partners^{7,8}, and potentially through the differential utilization of the sperm of individual males after insemination by more than one partner⁴. Although females can actively solicit and resist copulation, males can often overcome female mate choice by coercing females to copulate⁹ and sometimes manipulating them to preferentially use their sperm through specific copulatory structures and behaviours^{2–5}. Under these circumstances, cryptic female choice may be adaptive to females^{4,6}. Here we show that in the absence of specific male copulatory behaviours, female sperm choice based on one male phenotypic trait, social status, consistently explains female sperm ejection in the feral fowl.

Female fowl, *Gallus gallus domesticus*, may gain both direct and indirect fitness benefits by preferentially copulating with dominant males. High-ranking males provide better courtship feeding, anti-predator vigilance and protection from the sexual harassment of other males¹⁰. High-ranking males can also provide indirect benefits: they father more offspring than subordinates¹¹ because they obtain more copulations and compete more effectively for access to females^{10,12}; and dominance is heritable¹³. However, subordinate males are not prevented from copulating. Because they are bigger than females, males can coerce females into copulating despite female resistance. Female fowl, like many other taxa^{2–5}, can eject semen following insemination^{14,15}. We therefore considered whether female fowl would bias insemination success in favour of dominant males. We tested the prediction that sperm ejection was non-random with respect to male social status.

Females favoured inseminations by high-ranking males in two ways. First, they preferentially solicited copulations from high-ranking males (probability of solicitation versus male status, $r_s = 0.73$, $n_{\text{males}} = 13$, $P = 0.005$). Consistent with this general trend, individual females were more likely to solicit copulations from males of the top rather than the bottom half of the social hierarchy (mean ± standard error (s.e.) of probability of solicitation from high- and low-ranking males: 0.36 ± 0.06 versus 0.00 ± 0.00 , Wilcoxon $T = 0$, $P = 0.007$, $n_{\text{females}} = 10$). Second, females resisted most (71%, 534/747) copulations (range of resistance by individual females: 50–100%, mean ± s.e. percentage of resistant copulations per female = 49.52 ± 6.0), particularly those by low-ranking males (probability of resistance versus male status,

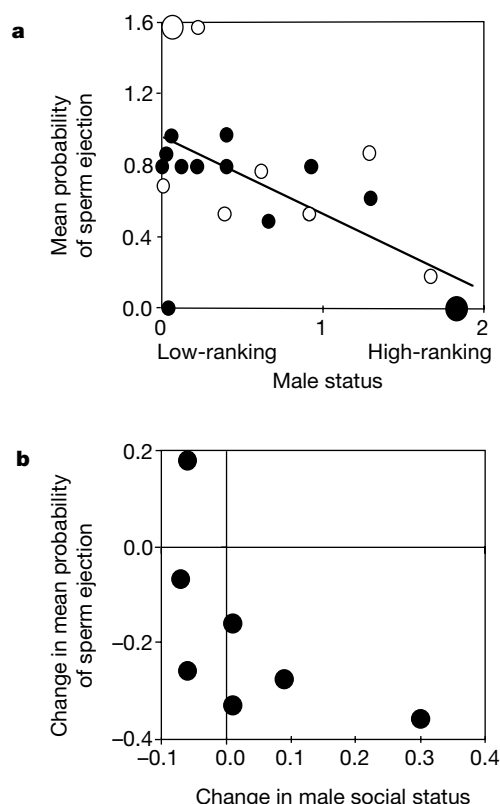


Figure 1 Male status and sperm ejections. **a**, Relationship between mean probability of sperm ejection (arcsine-transformed) and male social status (log-transformed) in 1998 (open circles) and in 1999 (filled circles). Small data points represent one male, large data points, two males. Effect on probability of sperm ejection: status: mean squares = 1.47, $F_{1,21} = 11.94$, $P = 0.003$; year: mean squares = 0.51 $F_{1,21} = 4.10$, $P = 0.06$, Sperm ejection = -0.43 ± 0.13 status + 0.98 ± 0.11 , $R^2 = 0.31$, $P = 0.004$, $n_{\text{copulations}}$: 1998 = 87, 1999 = 97. **b**, Relationship between the change in both the mean probability of sperm ejection and the male social status following the removal of six males ($r_s = 0.89$, $P = 0.007$, $n_{\text{males}} = 7$).

$r_s = -0.54$, $n_{\text{males}} = 13$, $P = 0.057$). Females were also more likely to utter a distress call when mounted by low-ranking males (mean $\pm$ s.e. male rank with and without distress call: 0.5 ± 0.07 versus 1.10 ± 0.1 , Wilcoxon $T = 18$, $n_{\text{females}} = 18$, $P = 0.003$), which increased the probability that dominant males would disrupt the copulation (mean $\pm$ s.e. probability of disruption with and without distress call: 0.86 ± 0.03 versus 0.34 ± 0.07 , Wilcoxon $T = 1$, $n_{\text{females}} = 18$, $P = 0.0003$, all males considered), which in turn reduced the probability of that copulation being successful (mean $\pm$ s.e. probability of success with and without disruption: 0.23 ± 0.02 versus 0.78 ± 0.04 , Wilcoxon $T = 0$, $n_{\text{females}} = 21$, $P = 0.001$, all males considered).

Despite these female pre-copulatory behaviours, low-ranking males were still able successfully to inseminate females. The probability of copulation success for the six lowest-ranking males was mean $\pm$ s.e. ($n_{\text{copulation partners}}$): 0.27 ± 0.05 (18), 0.22 ± 0.05 (17), 0.35 ± 0.06 (20), 0.24 ± 0.10 (14), 0.18 ± 0.06 (21), 0.29 ± 0.18 (17). In fact, male copulation competence was not dependent on male status. Considering only unresisted copulations in the absence of disruption, we found no difference in success between high- and low-ranking males (male status $F_{1,40} = 0.326$, $P = 0.57$, individual female effect $F_{14,40} = 0.97$, $P = 0.50$, $n_{\text{males}} = 10$, mean probability of copulation success $\pm$ s.e. of top half and bottom half of hierarchy: 0.87 ± 0.06 versus 0.92 ± 0.08 , $U = 6.5$, $n_{\text{top}} = 4$, $n_{\text{bottom}} = 6$, $P = 0.19$). Moreover, considering only resisted, non-disrupted copulations, low-ranking males were also as successful as high-

ranking males (male status $F_{1,39} = 0.002$, $P = 0.97$, individual female effect: $F_{17,39} = 1.67$, $P = 0.13$, $n_{\text{males}} = 11$, mean probability of copulation success $\pm$ s.e. of top half and bottom half of hierarchy: 0.50 ± 0.16 versus 0.28 ± 0.16 , $U = 8.0$, $P = 0.19$, $n_{\text{top}} = 5$, $n_{\text{bottom}} = 6$), indicating that, regardless of male status, female resistance had only a limited effect on copulation success.

Copulation consisted of the male mounting the female from the rear and the female raising her tail and everting her cloaca. Semen was ejaculated onto the cloaca as the male and female genitalia made contact for <1 s, after which the male dismounted ($n_{\text{copulations}} = 1,230$, see also ref. 16). Semen ejection from the cloaca occurred immediately after insemination and after the genitalia were disconnected, but while the male was still on the female's back or as he dismounted her. As a result, males could not and did not try to detect whether semen ejection had occurred (compare ref. 2), nor did male copulatory behaviour differ when semen ejection actually occurred.

Females were less likely to eject sperm of dominant males. We found a negative correlation between male status and the probability of sperm ejection per male (Fig. 1a). To analyse this in more detail, we examined data from individual females that copulated with ≥ 5 males and found that the relationship between male status and the probability of sperm ejection was more likely to be negative ($r_s < 0$ in 8/9 (89%) females, binomial test $P = 0.039$). Individual females were more likely to eject the sperm of low-ranking partners (combined probabilities from individual status versus probability of sperm ejection¹⁷, $\chi^2_{18} = 35.6$, $n_{\text{females}} = 9$, $P < 0.01$). Females that copulated with males from both the top and the bottom half of the social hierarchy also had a higher probability of ejecting sperm of low- rather than high-ranking males (mean $\pm$ s.e. probability of ejection of high- and low-ranking males: 0.76 ± 0.08 versus 0.37 ± 0.05 , Wilcoxon $T = 4$, $P = 0.028$, $n_{\text{females}} = 9$).

To explore this further, we manipulated the hierarchy by removing six males, including those occupying the top and the bottom rank. We found a negative correlation between the change in status and the change in the probability of sperm ejection (Fig. 1b). This result was consistent with the idea that females adjust their ejection behaviour according to male status. This experiment also led to a reduction in the probability of sperm ejection in all males but one, indicating that the decrease in the variance of male competitiveness caused by the removal of the top- and bottom-ranking males may have reduced female choosiness. This result cannot be explained by a decrease in the number of potential partners available to each female because we also removed 10 females to maintain female-to-male ratio.

We measured the volume of ejected ejaculates for these males and compared it with the mean volume of their natural ejaculate. The volume of an ejected ejaculate was within the range of variation of the volume of their natural ejaculates, suggesting that in most cases at least half of the ejaculate was ejected. The volume (in μl) of ejected ejaculate for the three males were: male 1: 25, (natural: 43.2, range: 22–97 ($n = 7$)), male 2: 45, (natural: 24.1, range: 7–87 ($n = 10$)), male 3: 60, (natural: 68.3, range: 43–82 ($n = 3$)).

A potentially confounding factor in the analysis of the effect of male status on the probability of sperm ejection is sperm depletion: because dominant males copulate more frequently than subordinates^{10,12}, ejaculate size may be constrained by sperm depletion¹⁸ and dominant males may produce smaller ejaculates than subordinates. If dominant males typically produce smaller ejaculates, this may have reduced the likelihood of sperm ejection or our ability to detect it. However, two lines of evidence showed that this was not the case. First, there was no difference in ejaculate volume (in μl) or the number of sperm in the ejaculate between the first (mean $\pm$ s.e. volume: 0.03 ± 0.003 , mean ($\times 10^6$) $\pm$ s.e. sperm number: 0.49 ± 0.096) and the second ejaculate produced within a 3-h period (volume: 0.03 ± 0.007 , sperm number: 0.43 ± 0.025) by a male (volume: Wilcoxon $T = 17.0$, $P = 0.28$,

sperm number: Wilcoxon $T = 9.5$, $P = 0.83$, $n_{\text{males}} = 10$). Second, and more importantly, we found a positive correlation between mean ejaculate volume and male status ($r_s = 0.68$, $n_{\text{males}} = 10$, $P = 0.03$).

When female pre-copulatory behaviour cannot prevent copulations from suboptimal partners, females can reduce the fertilization success of an ejaculate by increasing the intensity of sperm competition¹⁹ and by the differential ejection of sperm^{3,5}. This is the first study to show that sperm ejection in the absence of any form of male manipulation is non-random and that male phenotype can consistently explain variations in sperm ejection. Although females of some species may select against the sperm of incompatible genotypes^{6,20}, incompatibility avoidance is unlikely to generate directional selection because it is usually inconsistent between females²¹. Artificial insemination studies investigating cryptic female choice⁴ in fowl and other birds have failed to find evidence for incompatibility between the female reproductive tract or the ova and the sperm of individual males^{22,23}. However, our results indicate that during natural copulation cryptic female choice may occur through a behavioural response before sperm is taken up into the sperm storage tubules, rather than through a physiological response to sperm, and, in addition, that it may be consistent between females and have a directional character.

Differential sperm ejection may allow female birds to bias sperm utilization in favour of a preferred male phenotype. Because the number of sperm reaching the sperm stores of the female is crucial in determining fertilization success in birds²⁴, differential sperm ejection is likely to have a strong impact on the outcome of sperm competition in this group and in other taxa^{3,5}. Differential sperm ejection may allow females to influence paternity despite sexual coercion, and simultaneously reduce any costs associated with insemination and sperm storage^{25,26}. □

Methods

We studied a free-living population of feral fowl of the Bankiva breed, which is morphologically and behaviourally similar to the red junglefowl, *Gallus gallus gallus*^{10,27}. The investigation was conducted at the University of Stockholm from April to July 1998 when the study population consisted of 13 males and 21 females, and from April to July 1999 when the population consisted of 10 males and 13 females different from the previous year. Male hierarchy was established from the outcome of pairwise interactions²⁸ (male top rank = 1.83, bottom rank = 0.00). The birds were fully habituated and most copulations were observed from a distance ≤ 5 m. Solicited copulations were initiated by female crouching. In coerced copulations, a male mounted a female without her solicitation and this was followed by female resistance¹⁰. The probability of solicitation was measured as the proportion of copulations that were solicited by the female. The occurrence of sperm ejection was determined by observing whether an ejaculate was taken up or rejected through cloacal contractions within 5 s after insemination. Copulations where the ejaculate was not observed were not considered. The proportion of behaviourally successful copulations for which we could ascertain whether sperm ejection had occurred was 25.4% (87/343) in 1998. We collected natural ejaculates by allowing males to copulate with females provided with a false cloaca²⁹. We collected semen with a Gilson pipette and measured semen volume (± 0.5 – 1.0 μ l) and the number of sperm using a standard method³⁰. A general linear model was used to assess the importance of male status explaining the variation in the between-male probability of sperm ejection. The difference in social status and the mean probability of sperm ejection after the removal of six males were obtained as: $d(\text{status}) = \text{social status after removal} - \text{status before removal}$; $d(\text{ejection}) = \text{mean sperm ejection probability after removal} - \text{mean probability before removal}$.

Received 6 March; accepted 20 April 2000.

- Parker, G. A. Sperm competition and its evolutionary consequences in the insects. *Biol. Rev.* **45**, 525–567 (1970).
- Davies, N. B. Polyandry, cloaca-pecking and sperm competition in dunlocks. *Nature* **302**, 334–336 (1983).
- Otronen, M. Mating behavior and sperm competition in the fly, *Drosophila anilis*. *Behav. Ecol. Sociobiol.* **26**, 349–356 (1990).
- Eberhard, W. G. *Female Control: Sexual Selection by Cryptic Female Choice* (Princeton Univ. Press, Princeton, 1996).
- Edvardsson, M. & Arnqvist, G. Copulatory courtship and cryptic female choice in red flour beetles *Tribolium castaneum*. *Proc. R. Soc. Lond. B* **267**, 559–563 (2000).
- Birkhead, T. R. Cryptic female choice: criteria for establishing female choice of sperm. *Evolution* **52**, 1212–1218 (1998).
- Darwin, C. *The Descent of Man, and Selection in Relation to Sex* (Murray, London, 1871).
- Wiley, H. R. & Poston, J. Indirect mate choice, competition for mates, and coevolution of the sexes. *Evolution* **50**, 1371–1381 (1996).

- Clutton-Brock, T. H. & Parker, G. A. Sexual coercion in animal societies. *Anim. Behav.* **49**, 1345–1365 (1995).
- Pizzari, T. Reproductive strategies in feral fowl, *Gallus gallus*. Thesis, Univ. Sheffield (1999).
- Jones, M. J. J. & Mench, J. Behavioral correlates of male mating success in multisire flocks as determined by DNA fingerprinting. *Poultry Sci.* **70**, 1493–1498 (1991).
- Cheng, K. M. & Burns, J. T. Dominance relationship and mating behavior of domestic cocks—a model to study mate-guarding and sperm competition in birds. *Condor* **90**, 697–704 (1988).
- Craig, J. V., Ortman, L. L. & Guhl, A. M. Genetic selection for social dominance ability in chicks. *Anim. Behav.* **13**, 114–131 (1965).
- Allen, T. E. & Grigg, G. W. Sperm transport in the fowl. *Aust. J. Agr. Res.* **8**, 788–799 (1957).
- Thornhill, R. in *Sperm Competition in Birds: Evolutionary Causes and Consequences* (eds Birkhead, T. R. & Møller, A. P.) **82** (Academic, London, 1992).
- Etches, R. J. *Reproduction in Poultry* (CAB International, Oxon, 1996).
- Sokal, R. R. & Rohlf, J. F. *Biometry* 2nd edn (W. H. Freeman and Company, New York, 1981).
- Parker, J. E., McKenzie, F. F. & Kempster, H. L. Fertility in the male domestic fowl. *Univ. Miss. Agric. Exp. Station Bull.* **347**, 1–50 (1942).
- Simmons, L. W. Sperm competition as a mechanism of female choice in the field cricket *Gryllus bimaculatus*. *Behav. Ecol. Sociobiol.* **21**, 197–202 (1987).
- Clark, A. G., Begun, D. J. & Prout, T. Female $\times$ male interactions in *Drosophila* sperm competition. *Science* **283**, 217–220 (1999).
- Zeh, J. A. & Zeh, D. The evolution of polyandry II: post-copulatory defences against genetic incompatibility. *Proc. R. Soc. Lond. B* **264**, 69–75 (1997).
- Birkhead, T. R., Martinez, J. G., Burke, T. & Froman, D. Sperm mobility determines the outcome of sperm competition in the domestic fowl. *Proc. R. Soc. Lond. B* **266**, 1759–1764 (1999).
- Cunningham, E. J. A. & Cheng, K. M. Biases in sperm use in the mallard: no evidence for selection by females on sperm genotype. *Proc. R. Soc. Lond. B* **266**, 905–910 (1999).
- Martin, P. A., Reimers, T. J., Lodge, J. R. & Dziuk, P. J. The effect of ratios and numbers of spermatozoa mixed from two males on proportions of offspring. *J. Reprod. Fertil.* **39**, 251–258 (1974).
- Sheldon, B. C. Sexually transmitted disease in birds: occurrence and evolutionary significance. *Proc. R. Soc. Lond. B* **39**, 491–497 (1993).
- Chapman, T., Liddle, L. F., Kalb, J. M., Wolfner, M. F. & Partridge, L. Cost of mating in *Drosophila melanogaster* females is mediated by male accessory gland products. *Nature* **373**, 241–244 (1995).
- Harrison, B. Den Svenska dvärghönan. II. *Sven. Rasf. Tidskrift* **1**, 12–14 (1987).
- Clutton-Brock, T. H., Albon, S. D., Gibson, R. M. & Guinness, F. E. The logical stag: adaptive aspects of fighting in red deer (*Cervus elaphus* L.). *Anim. Behav.* **27**, 211–225 (1979).
- Ishikawa, H. The life duration of cock spermatozoa outside the body. *Proc. World Poultry. Fourth Congr.* **90**–94 (1930).
- Bakst, M. R. & Cecil, H. C. (eds) *Techniques for Semen Evaluation, Semen Storage, and Fertility Determination* (Poultry Sci. Assoc., Savoy, 1997).

Acknowledgements

We thank N. Andbjør, A. Bylin and C. K. Cornwallis for technical assistance, S. Jakobsson for providing facilities, A. P. Balmford, T. A. Burke, N. B. Davies, B. J. Hatchwell, F. M. Hunter, C. M. Lessells and P. J. Warren for comments. T.P. was supported by a Patrick & Irwin Packaging Fellowship and T.R.B. by a grant from the National Environmental Research Council.

Correspondence and requests for materials should be addressed to T.P. (e-mail: t.pizzari@sheffield.ac.uk).

A universal pattern of mortality decline in the G7 countries

Shripad Tuljapurkar, Nan Li & Carl Boe

Mountain View Research, 2251 Grant Road, Los Altos, California 94204, USA

Human lifespan has increased enormously this century^{1–3}. But we remain uncertain about the forces that reduce mortality, and about the cost implications of ageing populations^{4,5} and their associated social burden. The poor understanding of the factors driving mortality decline^{2,6,7}, and the difficulty of forecasting mortality^{8–11} are due in part to the pronounced irregularity of annual to decadal mortality change^{7,12}. Here we examine mortality over five decades in the G7 countries (Canada, France, Germany, Italy, Japan, UK, US). In every country over this period, mortality at each age has declined exponentially at a roughly constant rate. This trend places a constraint on any theory of society-driven mortality decline, and provides a basis for stochastic mortality forecasting. We find that median forecasts of life expectancy are substantially larger than in existing official forecasts. In terms of

the costs of ageing, we forecast values of the dependency ratio (that is, the ratio of people over 65 to working people) in 2050 that are between 6% (UK) and 40% (Japan) higher than official forecasts.

We examine long-term patterns in mortality rates, measured as central death rates $m(x, t)$ (that is, deaths in a year divided by mid-

year population) in age class x in year t for both sexes. We use annual data for the period 1950–1994 for all countries except Italy (1951–1993) and Germany (1952–1990); here, Germany means the former West Germany. We use the age classes, 0 to 1 yr, 1 to 5 yr, and then 5-yr intervals up to age 105; time t is in single years. For countries except Japan, ages 85 yr and older are reported as a single lumped class; for Japan, we used accurate national data for all age classes.

Our analysis echoes the pioneering work of Lee and Carter for the US¹³. We seek a dominant temporal ‘signal’ in the data—the simplest model that captures trend and variation in death rates—and to find it we use a singular-value decomposition^{13–15} applied to the logarithms $\log m(x, t)$ ^{13,16,17}. For each age x we subtract the sample average $a(x)$ of the logarithm and obtain the decomposition

$$\log m(x, t) - a(x) = \sum_i s_i u_i(x) v_i(t)$$

The real singular values are $s_1 \geq s_2 \geq \dots \geq 0$. The ratio of s_1^2 to the sum of squares of all singular values is the proportion of the total temporal variance in the transformed death rates that is explained by just the first term in the singular-value decomposition.

In every G7 country, the first singular value explains over 94% of the mortality variation (Table 1); therefore, we have a dominant temporal pattern, and we write

$$\log m(x, t) = a(x) + b(x)k(t) + E(x, t).$$

The single factor $k(t)$ corresponds to the dominant first singular value and captures most of the change in mortality. The far smaller variability from other singular values is $E(x, t)$. The dominant time factors $k(t)$ display highly linear long-term declines with superimposed short-term fluctuations (Fig. 1a). This linearity is supported by examination of differences (not shown) between consecutive values of $k(t)$; the term $E(x, t)$ has no long-term trend. Mortality decline at a particular age x depends on the profiles $b(x)$ (Fig. 1c): these are broadly similar but display differences attributable to social and historical factors¹⁸. We note that these age profiles will depend on the base period that we use for analysis: in comparing mortality change¹⁹ before and after 1950, the pattern of decline is similar but the earlier data reflect rapid declines in infant and child mortality.

What accounts for this regular long-term decline in mortality? An argument may be made if we assume that mortality decline in this century has resulted from a sustained application of resources and knowledge to public health and mortality reduction. Societies typically allocate attention and resources in proportion to observed levels of mortality at different ages (for example, immunization programs against childhood disease). Such allocation would produce an exponential (proportional) change in mortality, although not necessarily at a constant rate over time. Over time, the rate of proportional decline depends on a balance between the level of resources focused on mortality reduction and their marginal effectiveness. Historically, the level of resources has increased over time but their marginal effectiveness has decreased over time (because,

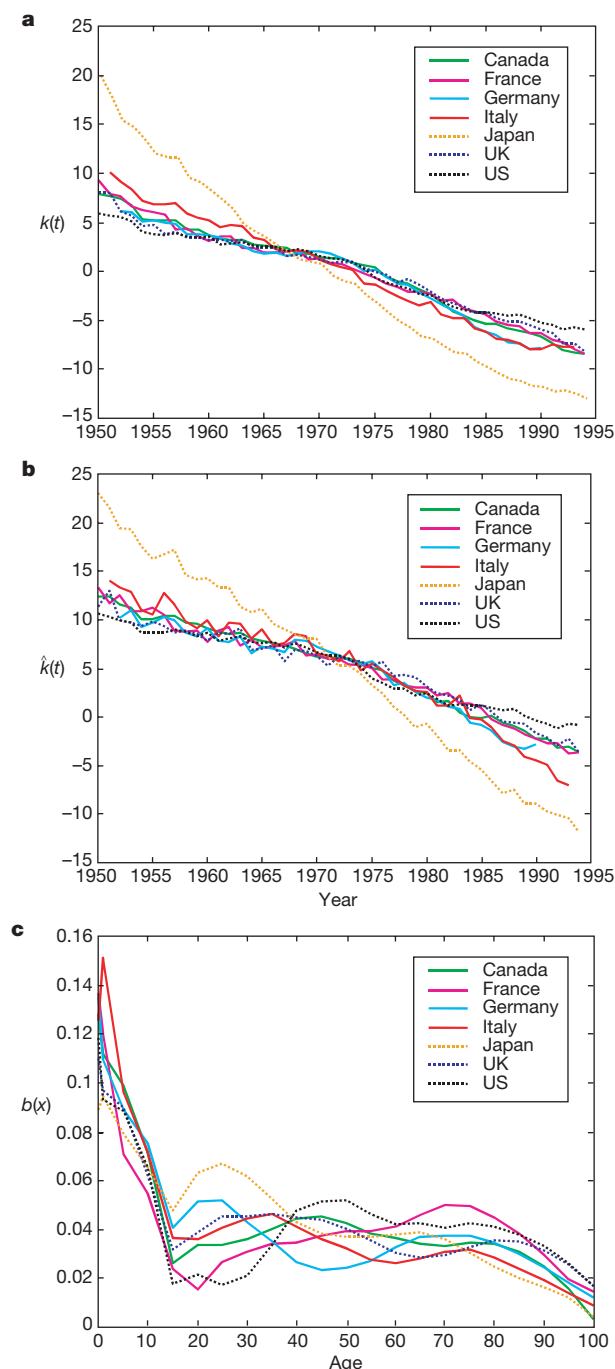


Figure 1 Dominant temporal patterns in the decline of mortality. **a, b**, Changes in age-specific central death rates are dominated by a single time factor $k(t)$, values of which are shown for the G7 countries (**a**). This time factor and its accompanying age factor, $b(x)$ account for over 94% of the temporal variation in the logarithm of the central death rates. To make forecasts that include variability not included in the dominant factor $k(t)$, we use a death-adjusted factor $\hat{k}(t)$ whose values are shown in **b, c**. The relative rate of progress against mortality at different ages. Mortality change at each age x depends on the product of the factor $k(t)$ and an age factor, $b(x)$, whose values are plotted versus age x for the G7 countries. The $b(x)$ for each country sum to one, so that the values shown indicate the relative rate of progress against mortality at different ages.

Table 1 Proportion of variance of $\log m(x, t)$ (both sexes combined) explained by the unadjusted mortality index $k(t)$ and the adjusted index $\hat{k}(t)$.

	Canada	France	Germany	Japan	Italy	UK	US
$k(t)$ (unadjusted)	0.9612	0.9428	0.9569	0.9520	0.9748	0.9479	0.9527
$\hat{k}(t)$ (adjusted)	0.9560	0.9323	0.9461	0.8912	0.9447	0.9321	0.9482

Table 2 The slope of adjusted mortality index $\hat{k}(t)$ and the value of in the adjusted model, for each country

	Canada	France	Germany	Italy	Japan	UK	US
Drift term (z)	0.3649	0.3877	0.3422	0.5043	0.7948	0.3427	0.2606
Standard error	0.3685	0.7017	0.7397	1.0367	0.7702	0.8765	0.3966

for example, we are confronted with ever more complex causes of mortality that require substantial resources or new knowledge). Our observation—that roughly constant long-run exponential rates of decline—implies that increasing level and decreasing effectiveness have balanced each other over long times. It is of course possible that the linear pattern of decline that we report has some other basis. For the future, we expect a continued increase in resources spent on mortality reduction and a growing complexity of causes of death. The balance between these could certainly shift if there were departures from history—for example, if new knowledge is discovered and translated into mortality reductions at an unprecedented rate. But this century has witnessed an amazing series of discoveries that have altered medicine and public health, and there is no compelling reason why the future should be qualitatively different. We therefore expect a continuation of the long-run historical pattern of mortality decline.

We use the long-term linear decline in $k(t)$ to forecast the

expectation of life at birth, e_0 , for the G7 countries. A naive forecast based on the long-run trend is not sensible because the short-term variation will accumulate over time. To incorporate short-term variation²⁰ we use a modified time factor defined by

$$\log m(x, t) \cong a(x) + b(x)\hat{k}(t)$$

and compute $\hat{k}(t)$ (Fig. 1b) to match historical total deaths in each year. This adjustment procedure is effective for the forecasts of life expectancy (e_0) we report here; the adjustment would have to be modified to forecast other quantities. In our view, differences in mortality trends between the sexes are mainly driven by short-term variations relative to aggregate mortality: for a sex-specific forecast we would model separately aggregate decline and the relative change between the sexes.

We use a stochastic model of decline,

$$\hat{k}(t+1) = \hat{k}(t) - z + \epsilon_t$$

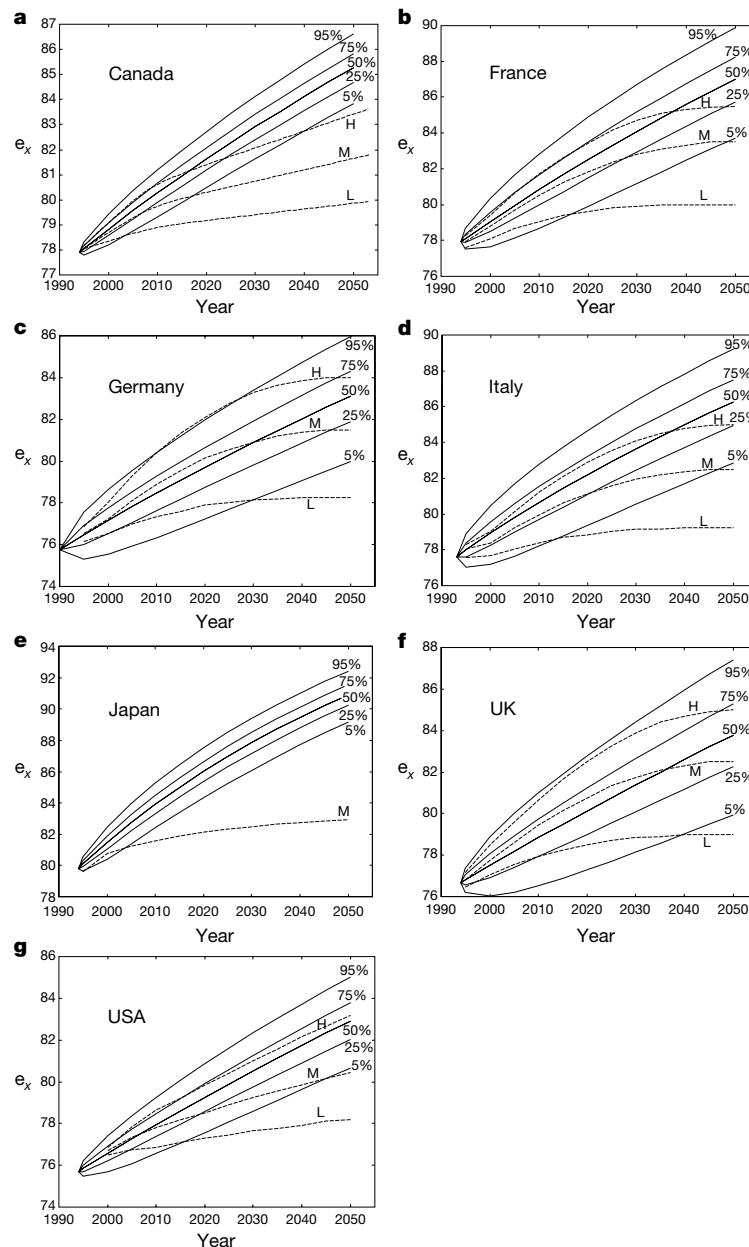


Figure 2 Forecasts of e_0 , the expectation of life corresponding to mortality rates for each year from 1995 to 2050. Stochastic (solid lines) and official (dashed lines) forecasts of e_0 are shown for the G7 countries. For the stochastic forecasts, point-wise prediction intervals at the 5%, 25%, 50%, 75% and 95% levels are shown; for the official forecasts,

high (H), medium (M) and low (L) outcomes are shown where they are available. Note the differences between stochastic and official forecasts in uncertainty over the span of the forecast and in the forecast levels.

Table 3 Stochastic forecasts of combined sex life expectancy at birth (e_0), compared with official forecasts

	Forecast	2000		2020		2050	
		Central	Range	Central	Range	Central	Range
Canada	S	78.83	1.24	81.64	2.23	85.26	2.78
	O	78.72	0.72	80.31	2.23	81.67	3.55
France	S	79.04	2.71	82.51	4.98	87.01	6.15
	O	78.80	1.30	81.85	3.85	83.50	5.50
Germany	S	77.16	3.07	79.71	4.75	83.12	5.97
	O	77.25	1.50	80.15	4.20	81.50	5.75
Italy	S	78.94	3.28	82.20	5.31	86.26	6.38
	O	78.40	1.35	81.15	4.05	82.50	5.75
Japan	S	81.47	2.07	86.01	3.20	90.91	3.27
	O	80.76	0.00	82.12	0.00	82.95	0.00
UK	S	77.50	2.85	80.13	5.50	83.79	7.47
	O	77.75	1.45	80.75	4.00	82.50	6.00
US	S	76.56	1.73	79.25	3.30	82.91	4.35
	O	76.70	0.35	78.50	2.55	80.45	5.00

Official forecasts have 'high', 'medium', and 'low' variants, except that there is only one official forecast for Japan²³. For France, Germany, Italy and the UK, mortality forecasts are those published by Eurostat (Data Shop Eurostat, Luxembourg)²⁴; For Canada, the forecasts follow assumptions of the Canadian Pension Plan^{25,26}. S, stochastic forecast: central (50%), range (95–5%). O, official forecast: central (medium), range (high–low).

where $z > 0$, and ϵ_t is a series of independent normal stochastic disturbances. This model has been used earlier to forecast mortality¹³ and population²¹ for the US. In countries other than Japan, age-specific mortality over age 85 is estimated using a model schedule^{7,17}.

The rate of decline z (Table 2, first row) is in the range of 0.3 to 0.5 per year in the US, Canada and the European countries, and is highest in Japan at about 0.8 per year. Note that exponential decline at a constant rate implies that the absolute annual decrement in mortality will diminish over time so that, for example, the rate of decline of life expectancy will also diminish over time. There is substantial short-term variability in the annual rates of decline (Table 2, second row).

Our forecasts advance from a launch date using the stochastic model to generate a probability distribution of forecast values in each year. Figure 2 exemplifies the difference between our stochastic forecasts of e_0 and official forecasts based on scenarios. For example, the median (50th percentile of probability) stochastic forecast for France is about 3.5 yr higher in 2050 than the central official forecast. Comparisons for all countries (Table 3) show stochastic median forecasts in 2050 that are higher than official ones by an astounding 8 yr for Japan, to a relatively low 1.3 yr for the UK. Official forecasts may be strongly influenced by short-term slow-downs in mortality change or by conservative expert opinion^{7,12}. To see why this matters, note that a 1-yr difference in e_0 corresponds to a difference of over 5% in the dependency ratio (population over 65 divided by population ages 20–64; we use the life-table ratio here, a good approximation for the G7). The dependency ratio is a useful direct index of the costs of ageing, so we project median dependency ratios, and thus costs of ageing, higher by between 6 and 40% than do official forecasts. Beyond 2050, this forecast divergence would increase, because virtually all official scenarios assume that mortality decline will eventually slow (or stop)²². Figure 2 and Table 3 also display differences in forecast uncertainty, which provides a measure of demographic and policy risk. Uncertainty grows much more quickly in the stochastic forecast, but by 2050, official 'high' and 'low' patterns spread over about the same range.

Our forecasts have large implications for the financing of public programs of old-age support and pensions, as well as private and public health insurance. In the US, for example, the finances of Social Security and medical care for the elderly will be driven by the ageing of the baby boom in the next two decades, but the longer term finances will turn on the sustained increase in human life span. The demographic forecasts that shape policy analyses related to old-age support should reflect the regularity of mortality change that we describe. □

Received 11 January; accepted 16 March 2000.

- Smith, D. W. *Human Longevity*. (Oxford Univ. Press, New York, 1993).
- Preston, S. H. in *Forecasting The Health of Elderly Populations* (eds Manton, K. G., Singer, B. H. & Suzman, R. M.) 51–77 (Springer, New York, 1991).
- Preston, S. H. *Mortality Patterns in National Populations, with Special Reference to Recorded Cause of Death*. (Academic, New York, 1976).
- Lee, R. D. in *Demography of Aging* (eds Martin, L. G. & Preston, S.H.) 8–49 (National Academy Press, Washington DC, 1994).
- World Bank Policy Research Report *Averting the Old Age Crisis: Policies to Protect the Old and Promote Growth* (Oxford Univ. Press, Oxford, 1994).
- McKeown, T. *The Modern Rise of Population* (Academic, New York, 1976).
- Tuljapourkar, S. & Boe, C. Mortality change and forecasting: How much and how little do we know? *N. Am. Actuarial J.* **2**, 13–47 (1998).
- Horiuchi, S. & Wilmoth, J. Age patterns of the life-table ageing rate for major causes of death: Aging, disease and mortality in Japan, 1951–1990. *J. Gerontol.* **A 52**, B67–B77 (1996).
- Wilmoth, J. R. Are mortality projections always more pessimistic when disaggregated by cause of death? *Math. Popul. Stud.* **5**, 293–319 (1995).
- Manton, K. G., Stallard, E. & Tolley, D. H. Limits to human life expectancy. *Popul. Dev. Rev.* **17**, 603–637 (1991).
- Meslé, F. Classifying causes of death according to an aetiological axis. *Popul. Stud.* **53**, 97–105 (1999).
- Wilmoth, J. R. in *Health and Mortality among Elderly Populations* (eds Caselli, G. & Lopez, A.) 266–287 (Oxford Univ. Press, Oxford, 1996).
- Lee, R. D. & Carter, L. Modeling and forecasting the time series of US mortality. *J. Am. Stat. Assoc.* **87**, 659–671 (1992).
- von Storch, H. Comment on statistics and physical oceanography. *Stat. Sci.* **9**, 215–221 (1994).
- Broomhead, D. & King, G. Extracting qualitative dynamics from experimental data. *Physica D* **20**, 217–236 (1986).
- de León, J. *Empirical EDA Models to Fit and Project Time Series of Age Specific Mortality Rates*. Technical Report No. 50. (Central Bureau of Statistics, Oslo, 1990).
- Himes, C. L., Preston, S. H. & Condran, G. A. A relational model of mortality at older ages in low mortality countries. *Popul. Stud.* **48**, 269–291 (1994).
- Lee, R. D. & Skinner, J. in *Assessing Knowledge of Retirement Behavior* (eds Hanushek, E. A. & Maritato, N. L.) 195–243 (National Academy Press, Washington DC, 1996).
- Wilmoth, J. R. The future of human longevity: a demographer's perspective. *Science* **280**, 395–397 (1998).
- Bell, W. Comparing and assessing time series methods for forecasting age-specific fertility and mortality rates. *J. Official Stat.* **13**, 279–303 (1997).
- Lee, R. D. & Tuljapourkar, S. Stochastic population forecasts for the United States: Beyond high, medium, and low. *J. Am. Stat. Assoc.* **89**, 1175–1189 (1994).
- Preston, S. H. Evaluation of postwar mortality projections in the United States, Canada, Australia, New Zealand and Japan. *World Health Statistics Report* **27**, 719–745 (1974).
- Population Projection of Japan: 1996–2050*. Technical Report (National Institute of Population and Social Security Research, Tokyo, 1997).
- de Beer, J. & de Jong, A. National population scenarios for countries of the European Economic Area. *Maandstatistiek van de Bevolking* **44**, 7–19 (1996).
- Goss, S., Wade, A., Bell, F. & Dussault, B. Historical and projected mortality for Mexico, Canada, and the United States. *N. Am. Actuarial J.* **2**, 108–126 (1998).
- Canada Pension Plan, Seventeenth Actuarial Report*, (Office of the Chief Actuary, Office of the Superintendent of Financial Institutions, Canada, 1998).

Acknowledgements

We thank M. Anderson, R. Lee and K. Wachter for comments. This work was supported by the US National Institute of Aging and the John Simon Guggenheim Foundation (S.T.).

Correspondence and requests for materials should be addressed to S.T. (e-mail: tulja@mvr.org).

Ultrasensitive pheromone detection by mammalian vomeronasal neurons

Trese Leinders-Zufall*, Andrew P. Lane*, Adam C. Puche*, Weidong Ma†, Milos V. Novotny†, Michael T. Shipley* & Frank Zufall*

* Department of Anatomy and Neurobiology and Program in Neuroscience, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA

† Department of Chemistry, Indiana University, Bloomington, Indiana 47405, USA

The vomeronasal organ (VNO) is a chemoreceptive organ that is thought to transduce pheromones into electrical responses that regulate sexual, hormonal and reproductive function in mammals^{1–5}. The characteristics of pheromone signal detection by vomeronasal neurons remain unclear^{3,5}. Here we use a mouse VNO slice preparation to show that six putative pheromones

evoked excitatory responses in single vomeronasal neurons, leading to action potential generation and elevated calcium entry. The detection threshold for some of these chemicals is remarkably low, near 10^{-11} M, placing these neurons among the most sensitive chemodetectors in mammals. Using confocal calcium imaging, we map the epithelial representation of the pheromones to show that each of the ligands activates a unique, nonoverlapping subset of vomeronasal neurons located in apical zones of the epithelium. These neurons show highly selective tuning properties and their tuning curves do not broaden with increasing concentrations of ligand, unlike those of receptor neurons in the main olfactory epithelium. These findings provide a basis for understanding chemical signals that regulate mammalian communication and sexual behaviour.

We investigated six structurally diverse ligands that have pheromonal activity in recipient female mice: 2,5-dimethylpyrazine, 2-sec-butyl-4,5-dihydrothiazole, 2,3-dehydro-*exo*-brevicomin, a mixture of *E,E*- α -farnesene and *E*- β -farnesene (referred to as farnesene), 2-heptanone, and 6-hydroxy-6-methyl-3-heptanone (Table 1). Each ligand is involved in diverse chemosignalling functions including oestrus induction or synchronization and puberty delay or acceleration (Table 1)^{6–10}. First, we used an intact VNO preparation to record extracellular field potentials from the microvillous surface of the sensory epithelium. These field poten-

tials, the electro-vomeronasogram (EVG), register summed local activities of vomeronasal neurons (VNs) and provide a measure of electrical responses to the ligands. Brief pulse-like applications of farnesene evoked negative field potentials in a concentration-dependent manner in the VNO of female mice (Fig. 1a). Only very low concentrations of farnesene, ranging from 10^{-11} to 10^{-10} M, were required to evoke a threshold response (Fig. 1a). Similar responses, but with different thresholds (10^{-11} – 10^{-8} M), were evoked by the remaining ligands (Table 1).

Because all six compounds produced negative deflections in the EVG response, we inferred that they generate ionic currents causing membrane depolarization leading to impulse generation in single VNs. To test this, we developed a VNO slice preparation¹¹ and used infrared videomicroscopy and extracellular loose-patch recordings to register activity from optically identified VNs (Fig. 1b). The probability of evoking stimulus-induced responses from a randomly selected VN was low. Of 52 VNs that were challenged sequentially with each of the six ligands, only three exhibited a chemoresponse and none responded to more than one compound. The responses consisted of a transient increase in impulse frequency (Fig. 1c, d). This indicates that VNs may be sharply tuned to specific

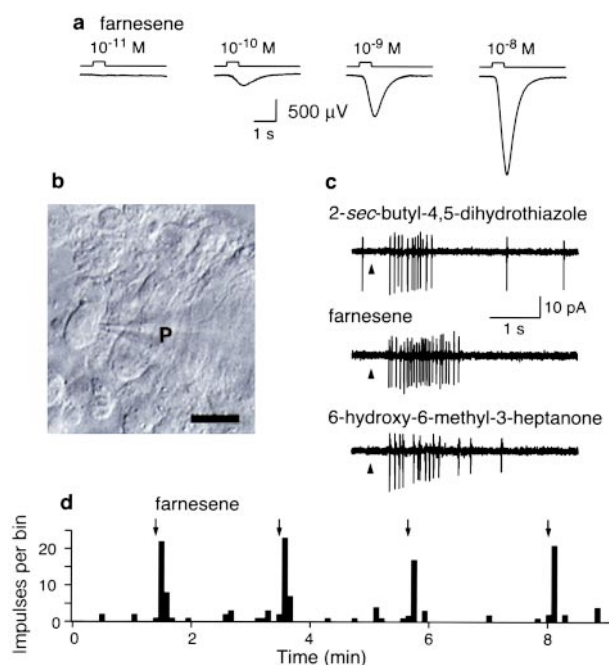


Figure 1 Pheromone-induced excitatory electrical responses in VNO. **a**, Negative field potentials in intact VNO and their dose-dependency produced by 500-ms pulses of increasing concentrations of farnesene. Ligands were focally applied to the microvillous surface of the sensory epithelium by a multi-barrelled stimulation pipette. **b**, Infrared-differential interference contrast micrograph showing VNO tissue slice. Individual VN somata are clearly visible. P, patch electrode. Scale bar, 10 μ m. **c**, Trains of extracellularly recorded, action-potential-driven, capacitive currents produced by bath application of 2-sec-butyl-4,5-dihydrothiazole, farnesene or 6-hydroxy-6-methyl-3-heptanone (all at 10^{-7} M). Each response was from a different VN using the approach shown in **b**. Pipette potential was 0 mV. Arrowheads, start of bath application (duration, 1–3 s). **d**, Analysis of the rate of spontaneous and stimulus-induced impulse discharges in a VN. Binwidth, 5 s. The same stimulus (farnesene, 10^{-7} M) was applied four times to show the repeatability of the responses. The level of spontaneous activity was low compared with stimulus-induced discharges. Although there is a slight decrease in the responses to stimulus three and four, it remains to be seen whether this is due to adaptation.

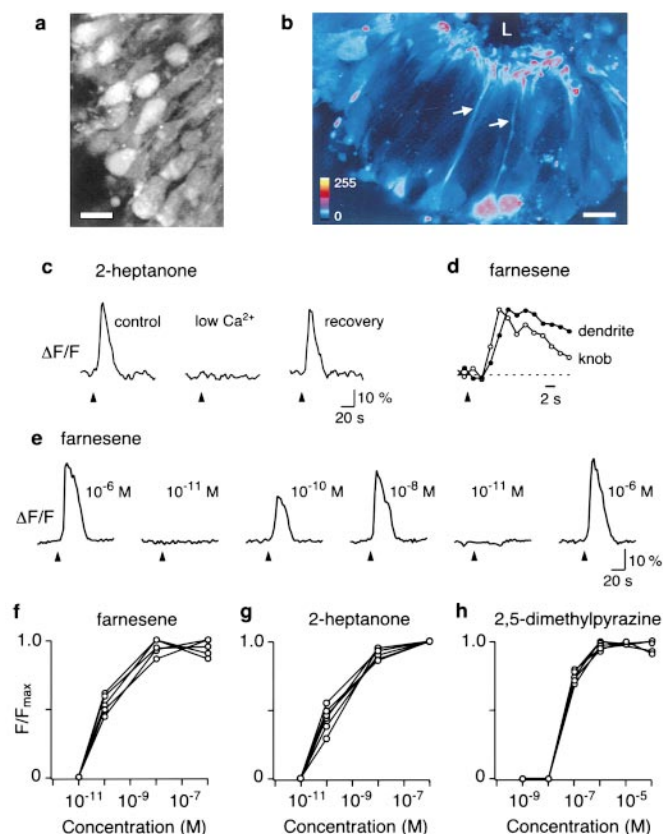


Figure 2 Pheromone-induced Ca^{2+} elevations in single VNs. **a**, Confocal fluorescence image (grey scale, acquired at rest) of a VNO slice loaded with fluo-4/AM showing ~ 50 VN somata. Scale bar, 10 μ m. **b**, Confocal fluorescence image acquired at rest (pseudocolour scale) showing subcellular structures of VNs such as single dendrites (arrows) projecting to the lumen (L) ending in knob-like swellings. Optical section was ~ 3 μ m thick. Scale bar, 10 μ m. **c**, Waveform of somatic Ca^{2+} transient of a single VN evoked by bath-applied 2-heptanone (10^{-7} M). The response was reversibly abolished by lowering the Ca^{2+} concentration in the bath solution to 0.6 μ M. **d**, Comparison of the time course of a farnesene-induced Ca^{2+} response (10^{-8} M) between knob dendrite of the same VN. **e**, Dose-dependency and reproducibility of farnesene-induced Ca^{2+} elevations measured in a single VN. **f–h**, Plots showing the dose-dependency of stimulus-induced peak Ca^{2+} responses of 17 different VNs. Each neuron responded either to farnesene, 2-heptanone or 2,5 dimethylpyrazine.

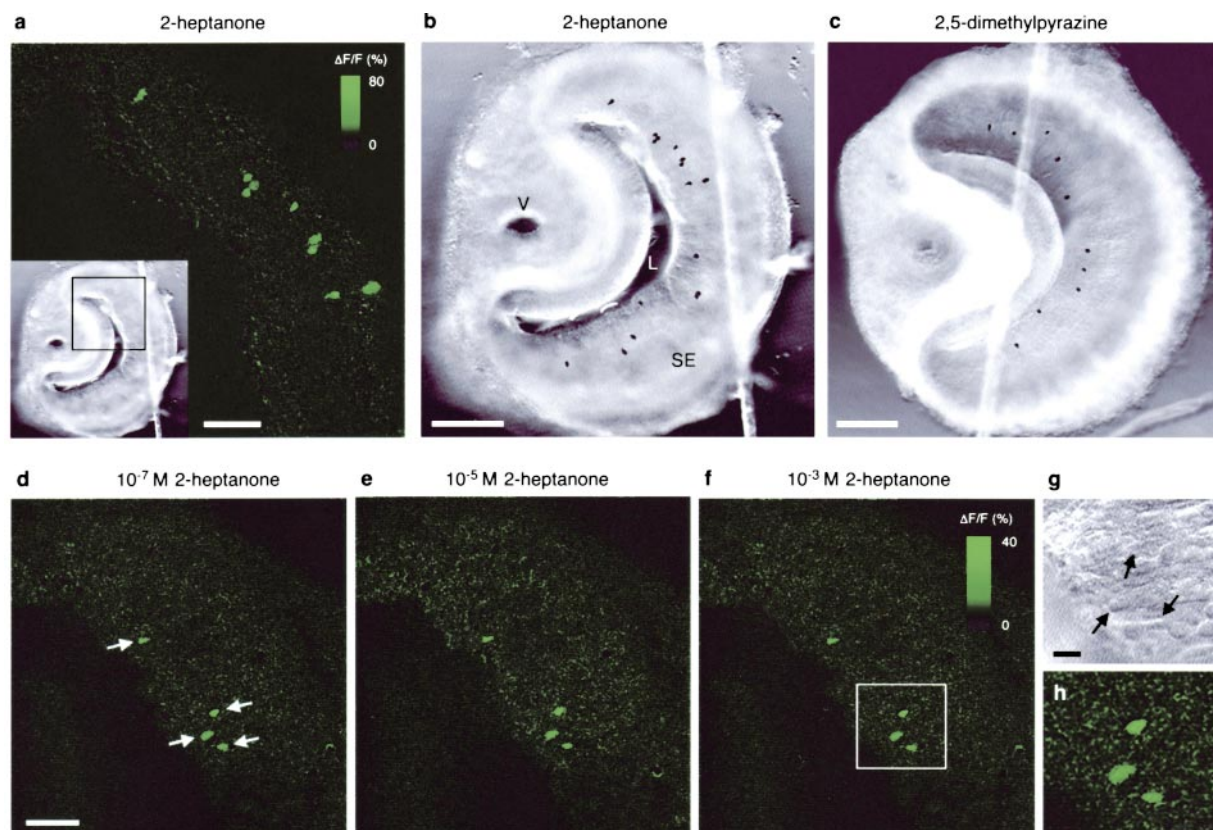


Figure 3 Spatial representation of functional, pheromone-induced activity in VNO. **a**, Pseudocolour image of the relative increase in stimulus-induced Ca^{2+} -dependent fluorescence (ratio between the peak fluorescence before and after stimulation, $\Delta F/F$) produced by bath-applied 2-heptanone (10^{-7} M), revealing nine activated VNs. Box in inset, position of the imaged region within the coronal VNO slice. Scale bar, 50 μm . **b**, **c**, Reconstructed VN response maps ($\Delta F/F$ Ca^{2+} images digitally superimposed onto a transmitted light image of the same slice) for 2-heptanone (10^{-7} M) and 2,5-dimethylpyrazine (10^{-7} M), showing that VNs activated by each of the two ligands

(black dots) were interspersed within the apical zone of the sensory epithelium (SE). V, vein. L, lumen. Scale bar, 100 μm . **d–f**, Concentration-independence of VN response maps. Successive stimulation using increasing concentrations of 2-heptanone (10^{-7} M, 10^{-5} M and 10^{-3} M) evoked an identical activity pattern consisting of four active VNs (arrows). Scale bar, 50 μm . **g**, the region delimited by the white box was plotted at higher magnification (transmitted light and fluorescence images) to show that the location of the fluorescence signals matched the location of single VNs. Scale bar, 10 μm .

pheromones such that only a small percentage of VNs respond to a given chemical. If so, recordings from one neuron at a time would be an impractical strategy for a systematic analysis of VN chemoresponsivity.

We therefore devised an optical imaging approach by which the responses of hundreds of VNs could be detected simultaneously. In olfactory receptor neurons (ORNs) of the main olfactory epithelium (MOE), odours induce a transient rise in intracellular Ca^{2+} , owing to the opening of Ca^{2+} -permeable transduction channels^{12–14}. Given that pheromones caused strong excitation, we hypothesized that VNs, too, would generate a Ca^{2+} elevation following chemostimulation. To test this, we loaded VNO slices with the Ca^{2+} indicator fluo-4/AM and used confocal microscopy to image them at cellular and subcellular resolution (Fig. 2a, b). Almost all VNs responded to extracellular application of KCl (60 mM) with a transient increase in Ca^{2+} (not shown). Spontaneous Ca^{2+} oscillations, which are often seen in developing VNs (results not shown), were not detected in older VNO slices. When we tested each of the six putative pheromones, all produced robust and reproducible increases in Ca^{2+} (measured at the soma) in a subset of VNs (Fig. 2c–g, see also Figs 3 and 4). The duration of these somatic Ca^{2+} responses were similar and, in some cases, shorter than odour-induced Ca^{2+} transients in ORNs^{13,14,24}. Removal of extracellular Ca^{2+} eliminated the responses in a reversible manner, providing evidence that they were triggered primarily by Ca^{2+} entry into the cells (Fig. 2c, $n = 3$). As in ORNs¹⁴, Ca^{2+}

responses were initiated at the distal end of the dendrite following chemostimulation, suggesting that they are a consequence of transduction events in the microvilli (Fig. 2d, $n = 6$).

To investigate whether the low thresholds observed in the EVG responses are due to unusually high sensitivity of individual VNs, we analysed stimulus–response relationships of single VNs (Fig. 2f–h). Each stimulus–response curve was derived from a single VN responding either to farnesene (Fig. 2f, $n = 5$), 2-heptanone (Fig. 2g, $n = 7$) or 2,5-dimethylpyrazine (Fig. 2h, $n = 5$). These curves allow several important inferences about the mechanisms of chemosensory coding in the VNO. First, single VNs have very low activation thresholds, near 10^{-11} M for farnesene and 2-heptanone, showing that VNs are remarkably sensitive to specific chemicals. The curves obtained with 2,5-dimethylpyrazine were further to the right, indicating that there are significant differences in sensitivity among VNs. Second, the responses saturate over about 2–3 orders of magnitude of ligand concentration, showing that VNs can encode stimulus intensity over a given concentration range. Third, VNs responding to the same chemical exhibit almost identical stimulus–response curves. Therefore, each of the cells responding to a given ligand may express a pheromone receptor with the same or very similar binding affinities.

To map the spatial organization of neurons responding to the same pheromone, we imaged larger regions of VNO slices while maintaining an optical resolution sufficient to detect responses in single VNs. Figure 3a shows a spatial pattern of neuronal activation

Table 1 Structure, origin and function of prospective pheromones used

Name	Chemical structure	Origin	Possible chemosignalling function in female mice	Detection threshold EVG response
2,5-dimethylpyrazine		Female urine	Puberty delay ⁶	10^{-8} – 10^{-7} M
2-sec-butyl-4,5-dihydrothiazole		Male bladder urine	Oestrus synchronization ⁷ Puberty acceleration ⁸	10^{-10} – 10^{-9} M
2,3-dehydro-exo-brevicomine		Male bladder urine	Oestrus synchronization ⁷ Puberty acceleration ⁸	10^{-10} – 10^{-9} M
α - and β -farnesenes		Male preputial gland	Puberty acceleration ⁸	10^{-11} – 10^{-10} M
2-heptanone		Female or male urine	Oestrus extension ⁹	10^{-11} – 10^{-10} M
6-hydroxy-6-methyl-3-heptanone		Male bladder urine	Puberty acceleration ¹⁰	10^{-8} – 10^{-7} M

EVG responses were measured as in Fig. 1a. Each detection threshold range is based on field potential measurements from at least three different mice at multiple locations on the VNO luminal surface.

following chemostimulation, revealing nine activated VNs within the imaged region. We then reconstructed 'VNO activation maps' for two of the chemicals, 2-heptanone and 2,5-dimethylpyrazine, by obtaining chemoresponse images for entire epithelial slices and digitally superimposing the responses onto a transmitted light

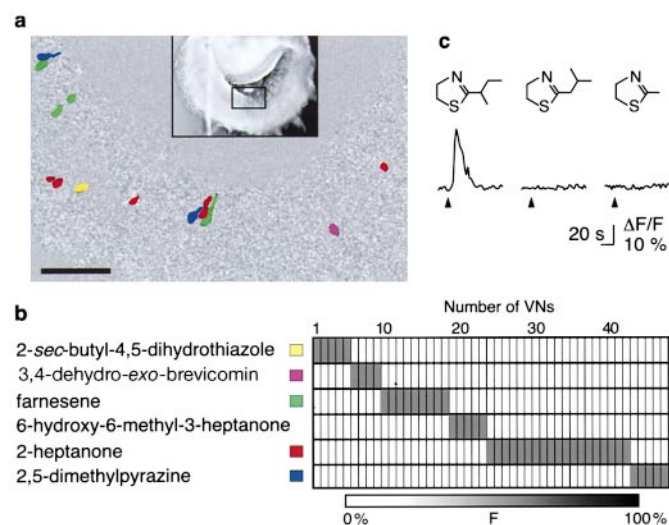


Figure 4 Chemoselectivity of VNs. **a**, VN activation map produced by successive stimulation with each of the six ligands listed in **b** (colour coded, each at 10^{-6} M). Each ligand activated a unique, nonoverlapping subset of VNs. Inset, low-power transmitted light image of the VNO slice; black box, imaged area. Scale bar, 50 μ m. **b**, Summary of the tuning profiles of 47 VNs (out of $\sim 1,600$ imaged VNs) responding to the six ligands. For each VN, the magnitude of the Ca^{2+} response was plotted as a percentage of the maximal response to a given chemical. Dark grey, 100%; white 0%. Without exception, VNs responded to only one of the pheromones tested. Thus, computed tuning curves were identical for all VNs that responded to the same pheromone. **c**, Two structural analogues of 2-sec-butyl-4,5-dihydrothiazole, isobutyl-4,5-dihydrothiazole and methyl-4,5-dihydrothiazole, were unable to evoke a Ca^{2+} response in VNs (each tested at 10^{-6} M). This result was confirmed with EVG recordings using 70-fold higher concentrations of the two analogues (not shown).

image of the same slice (Fig. 3b, c; $n = 7$). Three conclusions can be drawn from this analysis. First, a given chemical is detected by only a small percentage (0.2–3%) of all VNs (see also Fig. 4b). These values are strikingly similar to the percentage of cells that express the same putative pheromone receptor gene^{15–18}. Second, VNs activated by a given chemical are widely distributed in the epithelium with no significant difference between dorsal and ventral areas of the VNO. Analysis of several sections through the entire VNO revealed that activated VNs are randomly distributed along the anterior–posterior axis (not shown). This distribution of activated VNs is very similar to that previously seen for pheromone receptor gene expression^{15–18}. Third, VNs activated by any of the six chemicals were preferentially localized to the apical one-third of the sensory epithelium. Previous studies have shown that VNs in this apical zone express the G protein $\text{G}\alpha_2$ (refs 19, 20) and one family of putative pheromone receptors (V1Rs)¹⁵. Therefore, activation of VNs by the ligands we have examined may be restricted to $\text{G}\alpha_2$ - and V1R-positive cells.

Olfactory receptor neurons in the MOE are relatively broadly tuned to odours; each cell is sensitive to a different, but overlapping, set of odours and their tuning curves become less specific as odour concentration is increased. Thus, ORNs with the lowest threshold for a given odorant are activated first, and the less sensitive ones are recruited at higher concentrations, giving rise to a combinatorial coding scheme^{21–25}. We conducted two types of experiment to determine whether VNs employ such a combinatorial coding mechanism. First, we analysed the spatial patterns of VN activation as a function of stimulus concentration. Successive stimulations using increasing concentrations of 2-heptanone, from 10^{-7} M to 10^{-3} M, evoked an identical activity pattern consisting of four active VNs in the imaged region (Fig. 3d–j, $n = 3$). Other experiments using increasing doses of farnesene and 2,5-dimethylpyrazine gave the same result ($n = 5$), indicating that VNO activation maps are concentration-invariant.

Because a concentration-independent activation pattern may indicate that VNs use an unusually narrow tuning profile for stimulus discrimination, we next determined the chemoselectivity of VNs. Responses to successive stimulation with each of the six

compounds were mapped (Fig. 4a). Each pheromone activated an unique, non-overlapping, subset of VNs. We imaged around 1,600 VNs (in different slices from three mice) of which 47 cells responded to one of the six stimuli (Fig. 4b). In no case did a given VN respond to more than one of the six pheromones at any concentration tested. Sensitivity profiles were not only narrowly tuned but were identical for all VNs responding to the same pheromone (Fig. 4b). To investigate the tuning specificity of VNs further, we used two biologically inactive analogues⁶ of 2-sec-butyl-4,5-dihydrothiazole: isobutyl-4,5-dihydrothiazole and methyl-4,5-dihydrothiazole. None of the cells that were preferentially tuned for 2-sec-butyl-4,5-dihydrothiazole responded to the two analogues (Fig. 4c, $n = 7$). Furthermore, of about 1,300 imaged VNs with unknown tuning properties, none responded to the two analogues.

By combining electrophysiological and optical imaging techniques in VNO epithelial slices, we have systematically analysed responses to isolated pheromones in mammalian VNs. Our study establishes mouse VNs as ultrasensitive and highly selective pheromone detectors. The results indicate that the principles for processing of chemical information in the mammalian VNO may differ significantly from the combinatorial coding scheme used by the MOE^{21–25}. Mammalian VNs appear to be tuned to recognize only one or perhaps very few pheromonal components, thus resembling the narrowly tuned pheromone-detecting neurons of insects²⁵. With the ability to visualize pheromonal functional activity in the VNO in combination with genetic manipulation^{27,28} we can begin to decipher the logic of mammalian pheromonal communication. □

Methods

Tissue preparation

Female 30–60-day-old CD-1 mice were used for all experiments. For the EVG recordings from intact VNO, animals were decapitated, following anaesthesia, and the vomer bone capsule surrounding the VNO opened. The cavernous tissue was then removed to gain access to the luminal surface of the sensory epithelium. The tissue was superfused continuously at 23 °C with oxygenated saline (95% O₂/5% CO₂; solution A) containing (in mM): 120 NaCl, 25 NaHCO₃, 5 KCl, 5 BES (N,N-bis[2-hydroxyethyl]-2-aminoethanesulphonic acid), 1 MgSO₄, 1 CaCl₂, 10 glucose. For all other experiments, the dissected VNO was embedded in agarose, placed in ice-cold solution A, and coronal slices (250 µm) were prepared using a vibratome. After sectioning, slices were kept in solution A at 23 °C for up to 5 hours. All chemicals were purchased from Sigma unless otherwise stated.

Electrophysiology

Field potentials were recorded using glass pipettes (resistance 1–3 MΩ; tips filled with solution B in 1.5% agar) that were connected by an Ag/AgCl wire to a differential amplifier (DP-301, Warner Instruments). Solution B consisted of (in mM): 145 NaCl, 5 KCl, 10 Hepes, 1 MgCl₂, 1 CaCl₂, adjusted to pH 7.3 (NaOH) and 300 mOsm (glucose). A second Ag/AgCl wire connected to an agar bridge served as an indifferent electrode. The output signal was digitized, low-pass-filtered (8-pole Bessel; corner frequency 60 Hz), and analysed on-line using Pulse software (Heka Electronics). Action-potential-driven capacitive currents were recorded using patch pipettes filled with solution B (seal resistance ~80 MΩ) connected to an EPC-9 patch clamp amplifier (sample rate 150 µs per point).

Confocal Ca²⁺ imaging

Intracellular Ca²⁺ was monitored with fluo-4/AM (10 µM, Molecular Probes) using a Bio-RAD MRC-1024 confocal laser system (on Olympus BX60 microscope; 10×, 20× or 40× water immersion objectives). Optical sections were ~9 µm thick unless otherwise stated. Images were acquired at 0.4 Hz (ref. 29) and analysed using NIH Image 1.63 and Igor pro software (Wavemetrics). An estimate of the intracellular Ca²⁺ concentration in VN somata using calibration methods described previously²⁹ revealed resting values of 30–120 nM. Maximal somatic fluorescence changes ($\Delta F/F$) caused by chemostimulation were ~50%, corresponding to a change in Ca²⁺ concentration of ~550 nM assuming a K_D value of 345 nM. Saturation would produce a relative increase of ~80–110%. These high values were usually not reached in our experiments, ruling out the possibility that dye saturation was a significant problem.

Chemostimulation

Pheromones were either bath-applied (slice experiments, laminar flow chamber, flow rate ~100 µl s⁻¹) or focally ejected (EVG recordings) using multi-barrelled stimulation pipettes^{14,29}. Interstimulus interval was 4 min unless otherwise noted. Pheromone concentrations refer to those in the bath reservoir or stimulation pipette. The concentrations at the site of action were probably lower by a factor of two to four because of local dilution.

Synthetic 4,5-dihydrothiazoles, 2,3-dehydro-*exo*-brevicomin, and 6-hydroxy-6-methyl-3-heptanone were prepared as described^{10,26,30}, and diluted to final concentrations using solution B. Farnesene (Bedoukian Research), 2-heptanone (Fluka) and 2,5-dimethylpyrazine (Aldrich) were prepared in dimethyl sulphoxide (DMSO) and diluted to the final concentration using solution B. Final DMSO concentrations (<0.1%) had no effect on impulse discharges or Ca²⁺ concentrations.

Received 24 January; accepted 5 May 2000.

- Halpern, M. The organization and function of the vomeronasal system. *Annu. Rev. Neurosci.* **10**, 325–362 (1987).
- Wysocki, C. J. & Meredith, M. in *Neurobiology of Taste and Smell* (eds Finger, T. E. & Silver, W. L.) 125–150 (Wiley & Sons, New York, 1987).
- Keverne, E. B. The vomeronasal organ. *Science* **286**, 716–723 (1999).
- Bargmann, C. I. Olfactory receptors, vomeronasal receptors, and the organization of olfactory information. *Cell* **90**, 585–587 (1997).
- Tirindelli, R., Mucignat-Caretta, C. & Ryba, N. J. P. Molecular aspects of pheromonal communication via the vomeronasal organ of mammals. *Trends Neurosci.* **21**, 482–486 (1998).
- Novotny, M. V., Jemiolo, B., Harvey, S., Wiesler, D. & Marchlewska-Koj, A. Adrenal-mediated endogenous metabolites inhibit puberty in female mice. *Science* **231**, 722–725 (1986).
- Jemiolo, B., Harvey, S. & Novotny, M. V. Promotion of the Whitten effect in female mice by synthetic analogs of male urinary constituents. *Proc. Natl. Acad. Sci. USA* **83**, 4576–4579 (1986).
- Novotny, M. V., Ma, W., Wiesler, D. & Zidek, L. Positive identification of the puberty-accelerating pheromone of the house mouse: the volatiles associating with the major urinary protein. *Proc. R. Soc. Lond. B* **266**, 2017–2022 (1999).
- Jemiolo, B., Andreolini, F., Xie, T.-M., Wiesler, D. & Novotny, M. V. Puberty-affecting synthetic analogs of urinary chemosignals in the house mouse, *Mus domesticus*. *Physiol. Behav.* **46**, 293–298 (1989).
- Novotny, M. V. *et al.* A unique urinary constituent, 6-hydroxy-6-methyl-3-heptanone, is a pheromone that accelerates puberty in female mice. *Chem. Biol.* **6**, 377–383 (1999).
- Inamura, K., Kashiwayanagi, M. & Kurihara, K. Inositol-1,4,5-triphosphate induces responses in receptor neurons in rat vomeronasal sensory slices. *Chem. Senses* **22**, 93–103 (1997).
- Restrepo, D., Okada, Y. & Teeter, J. H. Odorant-regulated Ca²⁺ gradients in rat olfactory neurons. *J. Gen. Physiol.* **102**, 907–924 (1993).
- Bozza, T. C. & Kauer, J. S. Odorant response properties of convergent olfactory receptor neurons. *J. Neurosci.* **18**, 4560–4569 (1998).
- Leinders-Zufall, T., Greer, C. A., Shepherd, G. M. & Zufall, F. Imaging odor-induced calcium transients in single olfactory cilia: specificity of activation and role in transduction. *J. Neurosci.* **18**, 5630–5639 (1998).
- Dulac, C. & Axel, R. A novel family of genes encoding putative pheromone receptors in mammals. *Cell* **83**, 195–206 (1995).
- Herrada, G. & Dulac, C. A novel family of putative pheromone receptors in mammals with a topographically organized and sexually dimorphic distribution. *Cell* **90**, 763–773 (1997).
- Matsunami, I. & Buck, L. B. A multigene family encoding a diverse array of putative pheromone receptors in mammals. *Cell* **90**, 775–784 (1997).
- Ryba, N. J. P. & Tirindelli, R. A new multigene family of putative pheromone receptors. *Neuron* **19**, 371–379 (1997).
- Jia, C. & Halpern, M. Subclasses of vomeronasal receptor neurons: differential expression of G protein (G_{α2} and G_{αo}) and segregated projections to the accessory olfactory bulb. *Brain Res.* **719**, 117–128 (1996).
- Berghard, A. & Buck, L. Sensory transduction in vomeronasal neurons: evidence for G_{αo}, G_{αq}, and adenylyl cyclase II as major components of a pheromone signaling cascade. *J. Neurosci.* **16**, 909–918 (1996).
- Kauer, J. S. Contributions of topography and parallel processing to odor coding in the vertebrate olfactory pathway. *Trends Neurosci.* **14**, 79–85 (1991).
- Mori, K. & Yoshihara, Y. Molecular recognition and olfactory processing in the mammalian olfactory system. *Prog. Neurobiol.* **45**, 585–620 (1995).
- Duchamp-Viret, P., Chaput, M. A. & Duchamp, A. Odor response properties of rat olfactory receptor neurons. *Science* **284**, 2171–2174 (1999).
- Malnic, B., Hirono, J., Sato, T. & Buck, L. B. Combinatorial receptor codes for odors. *Cell* **96**, 713–723 (1999).
- Hildebrand, J. G. & Shepherd, G. M. Mechanisms of olfactory discrimination: converging evidence for common principles across phyla. *Annu. Rev. Neurosci.* **20**, 595–631 (1997).
- Zidek, L., Joo, A., Ma, W., Miao, Z. & Novotny, M. V. 4,5-dihydrothiazoles exhibit no correlations between their puberty-acceleration and activity for soluble binding proteins in the house mouse. *Chem. Senses* (submitted).
- Rodriguez, I., Feinstein, P. & Mombaerts, P. Variable patterns of axonal projections of sensory neurons in the mouse vomeronasal system. *Cell* **97**, 199–208 (1999).
- Belluscio, L., Koentges, G., Axel, R. & Dulac, C. A map of pheromone receptor activation in the mammalian brain. *Cell* **97**, 209–220 (1999).
- Leinders-Zufall, T., Rand, M. N., Shepherd, G. M., Greer, C. A. & Zufall, F. Calcium entry through cyclic nucleotide-gated channels in individual cilia of olfactory receptor cells: spatiotemporal dynamics. *J. Neurosci.* **17**, 4136–4148 (1997).
- Novotny, M., Schwende, F. J., Wiesler, D., Jorgenson, J. W. & Carmack, M. Identification of a testosterone-dependent unique volatile constituent of male mouse urine: 7-*exo*-ethyl-5-methyl-6,8-dioxabicyclo[3.2.1]-3-octene. *Experientia* **40**, 217–219 (1984).

Acknowledgements

This study was supported by grants from the NIH to T.L.-Z., M.V.N., M.T.S. and E.Z., and by an intramural grant from the University of Maryland (E.Z.). A.P.L. was a recipient of an NIDCD training grant. We thank R. Bock for programming spike analysis software.

Correspondence and requests for materials should be addressed to E.Z. (e-mail: zufu001@umaryland.edu).

Excessive placental secretion of neurokinin B during the third trimester causes pre-eclampsia

N. M. Page*, R. J. Woods*, S. M. Gardiner†, K. Lomthaisong*, R. T. Gladwell*, D. J. Butlin*, I. T. Manyonda‡ & P. J. Lowry*

* School of Animal and Microbial Sciences, The University of Reading, Reading RG6 6AJ, UK

† School of Biomedical Sciences, The University of Nottingham, Nottingham NG7 2UH, UK

‡ Department of Obstetrics and Gynaecology, St George's Hospital, London SW17 0QT, UK

Pre-eclampsia is a principal cause of maternal morbidity and mortality, affecting 5–10% of first pregnancies worldwide. Manifestations include increased blood pressure, proteinuria, coagulopathy and peripheral and cerebral oedema. Although the aetiology and pathogenesis remain to be elucidated, the placenta is undoubtedly involved, as termination of pregnancy eradicates the disease. Here we have cloned a complementary DNA from human placental messenger RNA encoding a precursor protein of 121 amino acids which gives rise to a mature peptide identical to the neuropeptide neurokinin B (NKB)¹ of other mammalian species. In female rats, concentrations of NKB several-fold above that of an animal 20 days into pregnancy caused substantial pressor activity. In human pregnancy, the expression of NKB was confined to the outer syncytiotrophoblast of the placenta, significant concentrations of NKB could be detected in plasma as early as week 9, and plasma concentrations of NKB were grossly elevated in pregnancy-induced hypertension and pre-eclampsia. We conclude that elevated levels of NKB in early pregnancy may be an indicator of hypertension and pre-eclampsia, and that treatment with certain neurokinin receptor antagonists may be useful in alleviating the symptoms.

The mammalian tachykinins comprise substance P² and neurokinins A³ and B¹. Their biological actions include smooth muscle contraction, vasodilation, pain transmission, neurogenic inflammation and activation of the immune system⁴. They are normally restricted to nervous tissue and exert their effects peripherally by their release from nerve endings which activates the neurokinin receptors, NK1, NK2 and NK3. Each NK peptide shows preference for one receptor: substance P for NK1; NKA for NK2; and NKB for NK3. NK1 and NK2 induce hypotension, whereas NK3 receptors cause contraction of the portal vein⁵, venoconstriction of the mesenteric bed⁶, and increased heart rate⁷—all potentially hypertensive. Although substance P and NKA can be found outside the brain, NKB was undetectable in peripheral tissues examined in non-pregnant animals⁸, despite being the most potent at NK3 receptors and exerting equally potent effects as substance P on the cerebral vasculature⁹.

During pregnancy, the activation of NK3 receptors by NKB would reduce the large blood flow through the liver to satisfy the need of the uterus and placenta. Pregnancy involves increases in maternal cardiac output and blood flow to the uterus, and a reduction in vascular resistance within the placental bed¹⁰. Trophoblastic invasion of the intradecidual portion of the spiral arteries renders them dilated, flaccid and unresponsive to vasoconstrictor agents and is a key haemodynamic event. Defective trophoblastic invasion is a consistent feature in pre-eclampsia, in which the spiral arteries retain their musculo-elastic properties and responsiveness to vasoactive substances¹¹, resulting in trophoblastic ischaemia. The endothelial cell damage that occurs in pre-eclampsia¹² leads to a reduction in the levels of the vasodilator and platelet-aggregation inhibitor, prostacyclin, with an increase in vasoconstrictor substances¹³.

Previously, we found that the increased concentrations of placental corticotropin releasing factor that we had observed in pregnancy-induced hypertension¹⁴ did not appear to contribute to the pathology of pre-eclampsia. We therefore sought other vasoactive placental neuropeptides using mRNA fingerprinting¹⁵ and the human data bases^{16,17} and found nine matches that showed high similarity to the bovine NKB precursor. Cloning the full-length gene

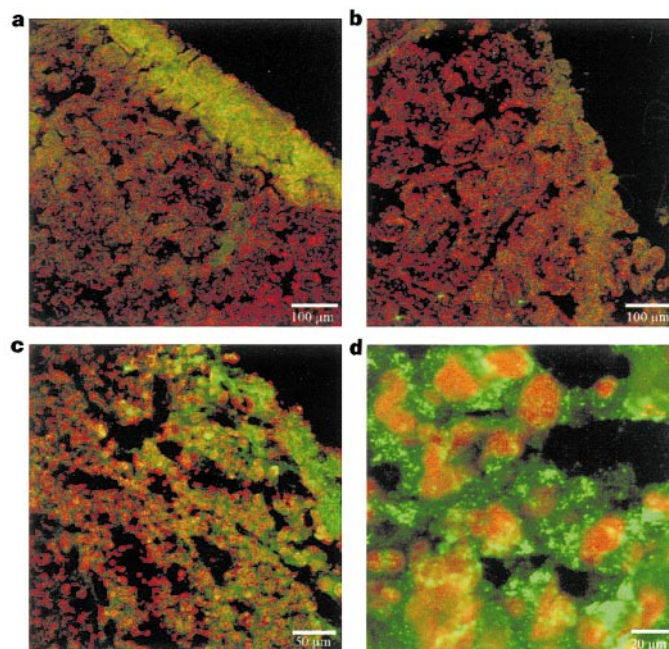


Figure 1 Localization of neurokinin B mRNA expression in vertical sections of the placenta. **a**, Human at term (39 weeks) with human antisense probe. **b**, Human at term (39 weeks) with human sense probe. **c**, Rat 18-day placenta with rat antisense probe.

d, High magnification showing giant cells of the rat placenta expressing neurokinin B. Magnification: **a**, 10× original size; **b**, 10×; **c**, 16×; **d**, 40×.

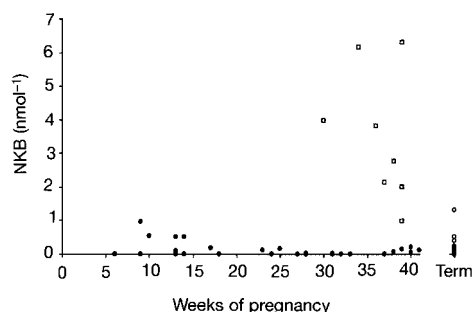


Figure 2 Neurokinin B levels in the plasma of 30 normotensive women (filled circles) were determined in samples taken throughout pregnancy and at term in 16 different women (open circles). They are compared with measurements made in eight pre-eclamptic subjects (open squares) in the last trimester of pregnancy. Pre-eclampsia was defined as follows: new hypertension (diastolic pressure consistently >90 mm Hg, with previous lower levels); and new proteinuria (>500 mg in 24 h) in the absence of urinary tract infection; both remitting remotely after delivery.

(GenBank accession number AF216586) revealed an open reading frame of 363 bases predicting the 121-residue human pro-neurokinin B (see Supplementary Information). The precursor protein gives rise to a mature peptide (DMHDFVGLM-NH₂) that is identical to neurokinin B of other mammalian species. Expression of placental NKB mRNA was detected in the first trimester and increased fivefold at term. *In situ* hybridization studies on the human placenta at term located it to the outer syncytiotrophoblasts ideally positioned to secrete its peptide into the maternal blood (Fig. 1a, b). Studies in rat placenta showed a similar expression of NKB mRNA in the junctional layer and in the giant cells (Fig. 1c, d).

Significant amounts of immunoreactive human NKB were found in early (21 pg g⁻¹) and term (25 pg g⁻¹) placentae; this NKB had chromatographic properties as assessed by high pressure liquid chromatography (HPLC) that were identical to synthetic NKB. An identical elution pattern was observed on extraction from term plasma. Samples taken from non-pregnant laboratory personnel all had low (near the sensitivity of the assay) or undetectable levels of the peptide, as did most samples taken from normotensive women throughout gestation, a proportion of whom exhibited slight increases at term (Fig. 2). Four individuals who had been admitted for elective abortions between weeks 9 and 14 had concentrations equivalent to the highest found at term (picomolar range); suggesting that the placentae from these individuals have started to secrete supraphysiological concentrations of NKB early in pregnancy. Patients in the third trimester suffering from pregnancy-induced hypertension and pre-eclampsia had concentrations in the nanomolar range (Fig. 2). Simultaneous samples taken from maternal and cord blood showed levels in cord blood that were roughly one-third of that of the mother (data not shown), indicating that hypersecretion of NKB could also affect NK receptors in the fetoplacental circulation. When NKB was undetectable in maternal blood, it was also undetectable in cord blood.

Samples taken from pregnant rats indicated that there is a rapid increase to nanomolar concentrations in maternal blood NKB on day 19, peaking at 3.1 nM on day 20, and then declining at term to 0.15 nM. In control female rats, concentrations were less than 50 pM. Intravenous infusion of low doses of NKB resulted in no significant change in blood pressure, but, during infusion of the highest dose (180 nmol h⁻¹) when the concentration of NKB reached was 20-fold higher (80 nM) than that found in the blood of a animal at 20-days pregnant, there was a significant, albeit transient, rise in mean arterial blood pressure (Fig. 3). Heart rate increase was not significant. The haematocrit value was $37.7 \pm 0.6\%$ in controls, and $39.0 \pm 1.1\%$, $36.0 \pm 0.7\%$ and

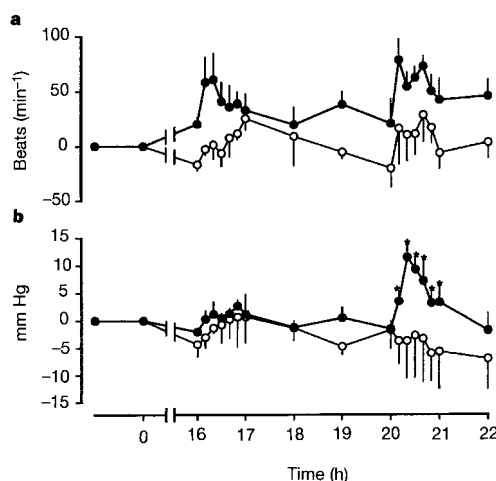


Figure 3 Changes in heart rate and blood pressure during infusion of saline (open circles, $n = 3$) or incremental doses of NKB (filled circles, $n = 6$) in conscious unrestrained female rats. **a**, Changes in heart rate; **b**, changes in blood pressure. NKB was infused at doses of 1.8 nmol h⁻¹ from $t = 0$; 18 nmol h⁻¹ from $t = 16$ h; and 180 nmol h⁻¹ from $t = 20$ h. Values are mean $\pm$ s.e.m.; asterisk, difference from the original baseline and from the values at $t = 20$ h is significant (Friedman's test).

$36 \pm 1.2\%$, respectively, after each of the three infusion periods (see Methods). In the animals receiving NKB there was a 37% gain in weight (0.624 ± 0.03 g compared with 0.459 ± 0.03 g, $P < 0.018$) of the uteri. Notably, infusion of NKB at the same doses in male rats had no effects on arterial blood pressure (data not shown).

These studies in conscious rats demonstrate a pressor effect of chronic high-dose infusion of NKB. We cannot determine the precise mechanism underlying the rise in blood pressure, but it is unlikely to have involved a change in blood volume, as control haematocrit values were not different from those at the end of the three infusion periods. There was, however, indirect evidence for an increase in blood to the uterus, as shown by a 37% increase in weight.

Previous reports have shown *in vivo* transient depressor¹⁸ or pressor¹⁹ effects of bolus injections of NK3 receptor agonists in anaesthetized and conscious rats, respectively. In the former study the effects were attributed to a Bezold-Jarisch reflex, and in the latter the changes were not mediated by NK3 receptors. In conscious guinea pigs, intravenous bolus doses of an NK3 receptor agonist did cause a transient rise in blood pressure which was NK3 receptor-mediated²⁰; however, to our knowledge we are the first to report cardiovascular effects of continuous infusion of NKB in conscious female animals to simulate late pregnancy concentrations. Although the effects were only significant at the highest dose, and then only transient, there are notable points. First, the effects occurred in conscious animals with fully competent cardiovascular reflexes. Second, endothelial release of nitric oxide acts to buffer the vasoconstricting effects of NKB, and, in pregnancy, there are a number of changes in the nitric oxide system²¹, which might influence the cardiovascular actions of the peptide. As the effects of NKB are relatively selective for the venous side of the mesenteric vascular bed, it is feasible that at low doses the only cardiovascular impact would be a modest rise in venous return and hence cardiac output. In this way, release of NKB from the normal placenta could contribute to the hyperdynamic circulation of normal pregnancy in the rat.

Could NKB be involved in the pathogenesis of pre-eclampsia? NKB could not cause the failed trophoblast invasion, but its potent and vasoactive properties might be involved in the clinical manifestations of pre-eclampsia. In response to the ischaemia consequent

upon defective trophoblastic invasion, the feto-placental unit must send a signal to the mother, and the obvious signal would be to maximize blood flow to the feto-placental unit by increasing the maternal blood pressure. Pre-eclampsia must be viewed as a continuum, because in mild cases of defective invasion there might be a small increase in blood pressure and none of the clinical features of the full-blown syndrome. When ischaemia within the feto-placental unit is profound, a stronger, enhanced and sustained signal is required. As NKB concentrations increase above normal, stimulation of the NK3 receptors causes constriction and contraction of the mesenteric and portal veins, increasing blood pressure, and potentially damaging the kidneys and liver. In severe cases, concentrations of NKB may be sufficient to stimulate peripheral NK1 receptors, for example, those found on platelets²² and neutrophils^{23,24}, and to contribute to other complications associated with activation of these cells. The NKB concentrations might also be responsible for the cerebral complications, as high intravascular concentrations of substance P have been shown to dilate cerebral blood vessels through NK1 receptors^{9,25}, which have been located to the endothelium²⁶. We found significant levels of NKB in the plasma of a small proportion of plasma samples from first trimester pregnancies; however, most were low throughout normotensive pregnancies and were increased in only a proportion at term. It is possible that increased secretion of NKB pre-dates the development of pre-eclampsia, and that increased levels of NKB in early pregnancy may identify pregnancies destined to develop the disease.

Potent and selective antagonists for the NK receptors may potentially be effective treatments for pre-eclampsia. NK3 receptor antagonists might alleviate the effects of high levels of NKB in the plasma of mothers suffering from hypertension, by reducing the vasoconstrictive effects of NKB on their mesenteric vascular beds and hepatic portal vein. Indeed, the potent NK3 receptor antagonist, GR138676, has been shown to be a competitive antagonist of contractions induced by the NK3 agonist, senktide, in the rat portal vein²⁷. Where concentrations of NKB have increased sufficiently to exert effects on NK1 receptors, for example, NK1 receptor antagonists might be the drugs of choice. Antagonists with a potency for NK receptors similar to that of NKB might offer better overall control because specific NK3 receptor antagonists, while alleviating the hypertensive effects, may induce a compensatory increase in placental NKB secretion and thus an increase in NK1 receptor agonist activity. □

Methods

Polymerase chain reaction with reverse transcription

Human placental tissue was obtained from pregnancy terminations at weeks 9 and 13, and at term, in compliance with and approval from the Local Research Ethics Committee. RNA was extracted with Tri Reagent (Sigma). The full-length preproNKB precursor was amplified from total RNA using SMART RACE cDNA amplification (Clontech, UK). 3' RACE was performed using a 5' gene-specific primer (5'-GGCACAGAGCTGCTCCACA GGCACCAT-3') derived from the human cDNA clone 138761 similar to bovine P08858 NKB precursor. The PCR fragment was purified following gel electrophoresis, cloned into the vector pGEM-T Easy (Promega), sequenced and compared with the GenBank database using the BLAST program.

In situ hybridization for NKB in the placenta

Sense and antisense digoxigenin (DIG) riboprobes were synthesized by *in vitro* transcription from rat or human NKB cDNA. Cryostat sections (25-μm) of 18-day rat placenta and 39-week human term placenta delivered by caesarian section were thaw-mounted on 3-aminopropyltriethoxysilane coated slides, fixed in paraformaldehyde and hybridized with the appropriate DIG-labelled riboprobe. Detection was performed using an anti-DIG antibody (Roche Diagnostics) coupled with the Renaissance Tyramide Signal Amplification System (New England Nuclear). Nuclei were counterstained with TOTO-3, and sections visualized using confocal microscopy.

Extraction and measurement of NKB from placenta and plasma

NKB was extracted from placenta as described for sheep brain²⁸ using similar weights of tissue to extraction medium. Extraction from plasma was performed as described for Met-enkephalin²⁹, but using octadecasilyl-silica cartridges, Sep-Pak Vac/ICC, C18 (Millipore Corporation). NKB eluted between 30 and 50% acetonitrile in aqueous 0.1% trifluor-

oacetic acid (v/v). Both extracts were reconstituted in 500 μl of NKB radioimmunoassay buffer (kit RIK 7357; Peninsula Labs, Belmont, CA), and assays carried out in duplicate according to manufacturer's protocols. Results were corrected by reference to extracted standards.

Cardiovascular effects of NKB in conscious rats

Female Sprague-Dawley rats (265–302 g; Charles River, UK) were anaesthetized with Hypnorm (0.126 mg kg⁻¹ fentanyl citrate, 4 mg kg⁻¹ fluanisone intraperitoneally (i.p.); Janssen, Belgium) and midazolam (5 mg kg⁻¹ i.p.; Antigen Pharmaceuticals, Ireland) for the implantation of intra-arterial (abdominal aorta via caudal artery) and intravenous (right jugular vein) catheters. After a 24-h recovery period, we measured resting arterial blood pressures and heart rates in the conscious unrestrained state for 2 h before continuous intravenous infusion of incremental doses of NKB (1.8 nmol h⁻¹ overnight for 16 h, 18 nmol h⁻¹ for 4 h, and 180 nmol h⁻¹ for 2 h; *n* = 6; Bachem, UK) or saline (0.4 ml h⁻¹; *n* = 3). Blood samples were measured before infusion of each dose for haematocrit and for plasma levels of NKB. Uteri and fallopian tubes were removed and weighed at the end. The animals were kept under constant lighting conditions for 48 h before and during experiments to ensure that they were in a similar state of oestrus.

Received 8 December 1999; accepted 4 April 2000.

- Kangawa, K., Minamino, N., Fukuda, A. & Matsuo, H. Neuromedin K: a novel mammalian tachykinin identified in porcine spinal cord. *Biochem. Biophys. Res. Commun.* **114**, 533–540 (1983).
- von Euler, U. S. & Gaddum, J. H. An unidentified depressor substance in certain tissue extracts. *J. Physiol.* **72**, 74–87 (1931).
- Kimura, S., Okada, K., Sugita, Y., Kanazawa, I. & Munekata, E. Novel peptides neuropeptides, neurokinin α and β, isolated from porcine spinal cord. *Proc. Japan Acad.* **B 59**, 101–104 (1983).
- Longmore, J., Hill, R. G. & Hargreaves, R. J. Neurokinin-receptor antagonists: pharmacological tools and therapeutic drugs. *Canadian J. Physiol. & Pharmacol.* **75**, 612–621 (1997).
- Mastrangelo, D. *et al.* The rat isolated portal vein: a preparation sensitive to neurokinins, particularly neurokinin B. *Eur. J. Pharmacol.* **134**, 321–326 (1987).
- D'Orleans-Juste, P., Claing, A., Telemaque, S., Warner, T. D. & Regoli, D. Neurokinins produce selective venoconstriction via NK-3 receptors in the rat mesenteric vascular bed. *Eur. J. Pharmacol.* **204**, 329–334 (1991).
- Thompson, G. W., Hoover, D. B., Ardell, J. L. & Armour, J. A. Canine intrinsic cardiac neurons involved in cardiac regulation possess NK1, NK2 and NK3 receptors. *Am. J. Physiol.-Regul. Integrative Comp. Physiol.* **44**, 1683–1689 (1998).
- Moussaoui, S. M. *et al.* Distribution of neurokinin B in rat spinal cord and peripheral tissues: comparison with neurokinin A and substance P and effects of neonatal capsicain treatment. *Neuroscience* **48**, 969–978 (1992).
- Jansen, I., Alafaci, C., McCulloch, J., Uddman, R. & Edvinsson, L. Tachykinins (substance P, neurokinin A, neuropeptide K, and neurokinin B) in the cerebral circulation: vasomotor responses *in vitro* and *in situ*. *J. Cereb. Blood Flow Metab.* **11**, 567–575 (1991).
- Baummann, H. & Huch, R. Maternal hemodynamics in pregnancy. *Am. J. Obstet. Gynecol.* **165**, 237–238 (1991).
- Lim, K. H. *et al.* Human cytotrophoblast differentiation/invasion is abnormal in pre-eclampsia. *Am. J. Pathol.* **151**, 1809–1818 (1997).
- Lyall, F. & Greer, I. A. The vascular endothelium in normal pregnancy and pre-eclampsia. *Rev. Reprod.* **1**, 107–116 (1996).
- Hayman, R., Brockelsby, J., Kenny, L. & Baker, P. Preeclampsia: the endothelium, circulating factor(s) and vascular endothelial growth factor. *J. Soc. Gynecol. Investig.* **6**, 3–10 (1999).
- Campbell, A. *et al.* Plasma corticotrophin-releasing hormone concentrations during pregnancy and parturition. *J. Clin. Endocrinol. Metab.* **64**, 1054–1059 (1987).
- Page, N. M., Manyonda, I., Butlin, D. J., Bicknell, A. B. & Lowry, P. J. Identification of novel differentially expressed secretory proteins from the human placenta. *J. Endocr.* **156** (Suppl.), P257 (1998).
- Boguski, M. S., Lowe, T. M. & Tolstoshev, C. M. dbEST—database for “expressed sequence tags”. *Nature Genet.* **4**, 332–333 (1993).
- Adams, M. D. *et al.* Initial assessment of human gene diversity and expression patterns based upon 83-million nucleotides of cDNA sequence. *Nature* **377**, Genome Directory 3S (1995).
- Couture, R., Laneuville, O., Guimond, C., Drapeau, G. & Regoli, D. Characterization of the peripheral actions of neurokinins and neurokinin receptor selective agonists on the rat cardiovascular system. *Naunyn-Schmied. Arch. Pharmacol.* **340**, 547–557 (1989).
- Cellier, E., Barbot, L., Regoli, D. & Couture, R. Cardiovascular and behavioural effects of intracerebroventricularly administered tachykinin NK3 receptor antagonists in the conscious rat. *Br. J. Pharmacol.* **122**, 643–654 (1997).
- Roccon, A., Marchionni, D. & Nisato, D. Study of SR 142801, a new potent non-peptide NK3 receptor antagonist on cardiovascular responses in conscious guinea-pig. *Brit. J. Pharmacol.* **118**, 1095–1102 (1996).
- Conrad, K. P., Kerchner, L. J. & Mosher, M. D. Plasma and 24h NOx and cGMP during normal pregnancy and preeclampsia in women on a reduced NOx diet. *Am. J. Physiol.* **46**, F48–F57 (1999).
- Gecse, A., Kis, B., Mezei, Z. & Telegdy, G. The effect of bradykinin and substance P on the arachidonate cascade of platelets. *Immunopharmacology* **33**, 167–170 (1996).
- Perianin, A., Snyderman, R. & Malfroy, B. Substance P primes human neutrophil activation: a mechanism for neurological regulation of inflammation. *Biochem. Biophys. Res. Commun.* **161**, 520–524 (1989).
- Perretti, M., Ahluwalia, A., Flower, R. J. & Manzini, S. Endogenous tachykinins play a role in IL-1-induced neutrophil accumulation: involvement of NK-1 receptors. *Immunology* **80**, 73–77 (1993).
- Kobari, M. *et al.* Intravascular substance P dilates cerebral parenchymal vessels through a specific tachykinin NK1 receptor in cats. *Eur. J. Pharmacol.* **317**, 269–274 (1996).
- Shimizu, T. *et al.* Occurrence and distribution of substance P receptors in the cerebral blood vessels of the rat. *Brain Res.* **830**, 372–378 (1999).
- Stables, J. M. *et al.* GR138676, a novel peptidic tachykinin antagonist which is potent at NK3 receptors. *Neuropeptides* **27**, 333–341 (1994).

28. Baigent, S. M. & Lowry, P. J. Urocortin is the principal ligand for the corticotrophin-releasing factor binding protein in the ovine brain with no evidence for a sauvagine-like peptide. *J. Mol. Endocrinol.* **24**, 53–63 (2000).
29. Clement-Jones, V., Lowry, P. J., Rees, L. H. & Besser, G. M. Met-enkephalin circulates in human plasma. *Nature* **283**, 295–297 (1980).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

This work was supported by the Medical Research Council.

Correspondence and requests for materials should be addressed to P.J.L. (e-mail: p.j.lowry@reading.ac.uk).

The γ -subunit of the coatmer complex binds Cdc42 to mediate transformation

Wen Jin Wu, Jon W. Erickson, Rui Lin & Richard A. Cerione

Departments of Molecular Medicine, and Chemistry and Chemical Biology, VMC, Cornell University, Ithaca, New York 14853-6401, USA

The Ras-related GTP-binding protein Cdc42 is implicated in a variety of biological activities including the establishment of cell polarity in yeast, the regulation of cell morphology, motility and cell-cycle progression in mammalian cells and the induction of malignant transformation^{1,2}. We identified a Cdc42 mutant (Cdc42F28L) which binds GTP in the absence of a guanine nucleotide exchange factor, but still hydrolyses GTP with a turnover number identical to that for wild-type Cdc42 (ref. 3). Expression of this mutant in NIH 3T3 fibroblasts causes cellular transformation, mimicking many of the characteristics of cells

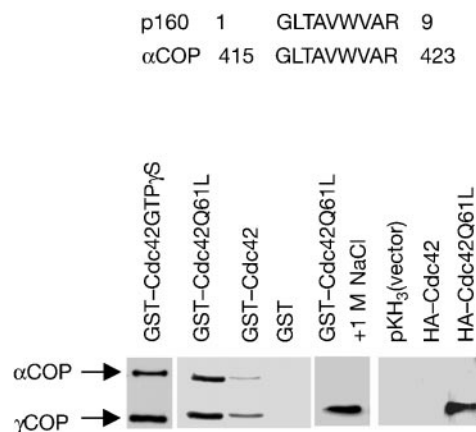


Figure 1 Coatmer subunits associate with activated Cdc42. Top, alignment of the nine amino acids sequenced from the M_r 160K putative Cdc42-target with residues 415–423 of the α COP subunit. Bottom, the first five lanes are the affinity precipitations of coatmer subunits (α COP and γ COP) from NIH 3T3 cell lysates using different GST-coupled Cdc42 fusion proteins. We loaded the GTPase-defective Cdc42 (Cdc42Q61L) with GTP and wild-type Cdc42 with GTP γ S or GDP. In lane 5, we washed the glutathione agarose beads in 1M NaCl (3 $\times$) to disassemble the coatmer subunits¹³. The last two lanes are HA-tagged Cdc42 proteins that immunoprecipitated from COS-7 cells transiently transfected with pKH3 constructs encoding the indicated Cdc42 proteins. We detected the α - and γ -coatmer subunits by western blotting with an antibody that recognizes both of these subunits.

transformed by the Dbl oncoprotein, a known guanine nucleotide exchange factor for Cdc42 (ref. 4). Here we searched for new Cdc42 targets in an effort to understand how Cdc42 mediates cellular transformation. We identified the γ -subunit of the coatmer complex (γ COP) as a specific binding partner for activated Cdc42. The binding of Cdc42 to γ COP is essential for a transforming signal distinct from those elicited by Ras.

A number of putative targets for Cdc42 have been identified, including the p21-activated kinases (PAKs), the Wiscott–Aldrich syndrome proteins (WASPs) and IQGAP (refs 5–8). These targets are involved in Cdc42-mediated changes in the actin cytoskeleton and cell morphology, but targets/ effectors that are responsible for Cdc42-regulated cell growth or malignant transformation have remained elusive. Thus, we generated glutathione S-transferase (GST)–Cdc42 fusion proteins to use as affinity resins to screen NIH 3T3 cell lysates for novel Cdc42 targets. In addition to previously identified binding partners of Cdc42, this screen revealed a protein with relative molecular mass of 160,000 (M_r 160K) which binds to activated, GTPase-defective GST–Cdc42Q61L, better than wild-type GDP-bound GST–Cdc42 or GST alone (not shown). We sequenced the M_r 160K protein band and found that it is the α COP subunit (Fig. 1), a component of the coatmer protein complex. This complex is important in intracellular trafficking because of its ability to assemble at Golgi membranes under the direction of the GTP-binding protein Arf (refs 9–11).

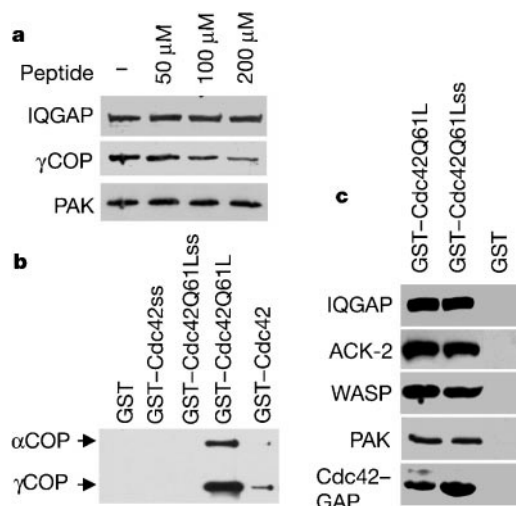


Figure 2 A dilysine sequence within the carboxy-terminal end of Cdc42 is necessary for binding coatmer. **a**, A synthetic peptide corresponding to the C-terminal tail sequence of cargo receptors blocks the binding of activated Cdc42 to γ COP. We incubated glutathione-agarose bound GST–Cdc42Q61L with NIH 3T3 cell lysates containing the indicated amounts of the synthetic peptide CYLRFFKAKKLE. We then precipitated the beads and assayed for the association of γ COP using the anti- α/γ COP antibody, IQGAP using an IQGAP antibody raised against the amino terminus²⁹ or PAK using an anti-PAK polyclonal antibody³⁰. **b**, Substitution of lysine residues 183 and 184 of Cdc42 to serines prevents the binding of Cdc42 to γ COP. We determined the association of coatmer subunits with GST–Cdc42, GTPase-defective Cdc42 (GST–Cdc42Q61L), and the corresponding GST–Cdc42 proteins that contain serine rather than lysine at positions 183 and 184 (GST–Cdc42ss and GST–Cdc42Q61Lss, respectively) as in Fig. 1. **c**, Lysine residues 183 and 184 are not essential for the binding of Cdc42 to other targets and regulators. We incubated the indicated GST–Cdc42 proteins or GST alone with lysates from NIH 3T3 cells that express endogenous IQGAP and PAK, lysates that ectopically expressed HA-tagged ACK-2 or Myc-tagged WASP, or with purified recombinant Cdc42GAP, and then precipitated and analysed for their binding by western blotting. IQGAP and PAK were detected using polyclonal antibodies^{29,30}; HA–ACK2 and Myc–WASP were detected with anti-HA and anti-Myc antibodies, and the recombinant purified Cdc42GAP was detected by Coomassie blue staining.

We obtained additional evidence that coatomer subunits bind to activated Cdc42 using an antibody that recognizes the α COP and γ COP subunits (883; a gift from C. Harter and F. Wieland)¹². When GST–Cdc42Q61L (containing bound GTP) was incubated with NIH 3T3 cell lysates, both α - and γ -coatomer subunits were affinity precipitated upon the addition of glutathione-agarose (Fig. 1). An antibody against the β COP subunit produced similar results (not shown), indicating that the entire coatomer complex associates with activated Cdc42. GST–Cdc42 (wild type) loaded with GTP γ S also forms a stable complex with the coatomer subunits, whereas Cdc42 bound to GDP is significantly less effective in binding coatomer and GST alone does not bind coatomer. Cdc42 appears to be unique among Rho subfamily proteins in its ability to interact with the γ COP subunit, as activated, GTPase-defective versions of Rac1 and RhoA showed little or no detectable binding to γ COP (not shown).

The endogenous coatomer subunits can be co-immunoprecipitated with haemagglutinin (HA)-tagged Cdc42Q61L that is transiently expressed in COS-7 cells, but not with HA–Cdc42 (wild-type protein) (Fig. 1), because the wild-type protein is predominantly in the GDP-bound state, which has weak affinity for targets. In this particular experiment, we detected the γ COP subunit but very little α COP. Although the relative amount of α COP detected in these precipitates varies, the stronger detection of γ COP appears to reflect its direct binding to Cdc42. When we incubated GST–Cdc42Q61L with NIH 3T3 cell lysates treated with 1 M NaCl, a condition that causes dissociation of the coatomer subunits¹³, only the γ COP subunit was affinity precipitated (Fig. 1, compare lanes 2 and 5). Moreover, purified recombinant GST– γ COP binds to purified recombinant Cdc42Q61L (not shown). These results indicate that activated Cdc42 binds the γ COP subunit directly and that other coatomer subunits (for example, α COP) associate with Cdc42 indirectly as part of the coatomer complex.

A number of lines of evidence show that the coatomer complex binds to retrieval receptors in the endoplasmic reticulum and to cargo receptors in the Golgi through a carboxyl-terminal dilysine motif^{12,14–16}. A peptide analogous to the cytoplasmic tail of cargo receptors, CYLRRFFKAKKLE, is sufficient for binding γ COP¹⁵. Addition of this peptide to a reaction containing GST–Cdc42Q61L and γ COP from cell lysates caused a dose-dependent inhibition of

the binding of γ COP to Cdc42 (Fig. 2a). The binding of GST–Cdc42Q61L to IQGAP, a major cellular target of Cdc42 (ref. 17) or to the serine/threonine kinase PAK3 (ref. 6), was completely unaffected by the peptide.

Cdc42 has consecutive lysines within its C-terminal region at positions 183 and 184. Mutating these lysine residues to serines eliminated the binding of the γ COP subunit to Cdc42 (Fig. 2b), without altering its ability to bind or hydrolyse GTP or bind to any

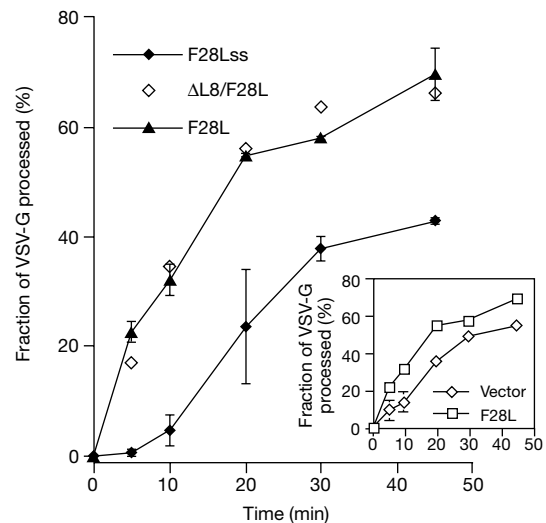


Figure 4 Comparisons of the effects of Cdc42F28L and Cdc42F28Lss on the rate of intracellular transport of the VSV-G protein. We used NIH 3T3 cells that stably express a constitutively active form of Cdc42F28L or a mutant Cdc42F28L in which lysine residues 183 and 184 were changed to serines (F28Lss). F28Lss cells have a reduced rate of VSV-G transport compared with F28L cells, indicated by their susceptibility to endoglycosidase digestion. Data are the average of two experiments. The rate of VSV-G transport in NIH 3T3 cells that express a constitutively active form of Cdc42 in which the insert region has been removed (Δ L8/F28L) is indistinguishable from that in F28L cells. The inset compares the rates for VSV-G transport in cells expressing Cdc42F28L with control cells transfected with vector alone.

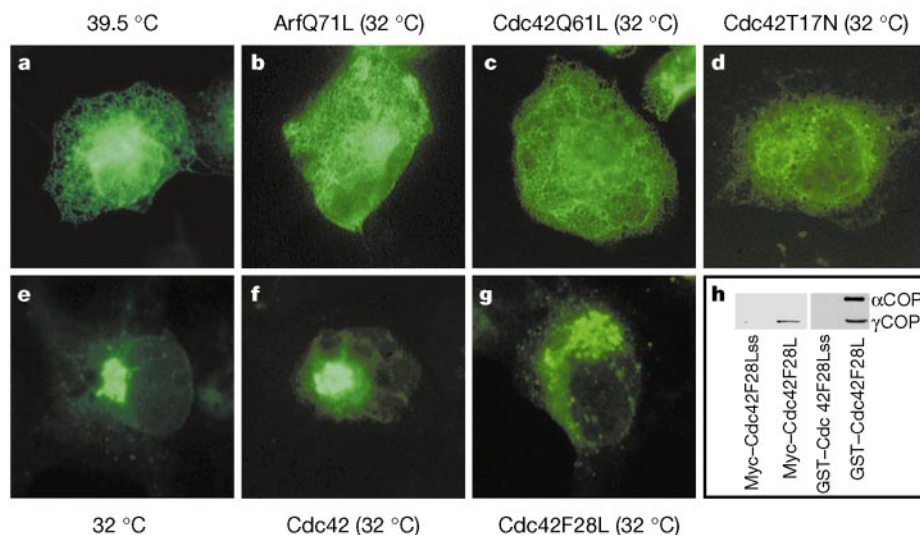


Figure 3 Effects of Cdc42 and Arf1 on the ER to Golgi transport of VSV-G protein. **a–g**, Indirect immunofluorescence of VSV-G, using the mouse monoclonal anti-VSV-G antibody (P5D4), for cells expressing the indicated forms of HA-tagged Cdc42 or HA-tagged (GTPase-defective) Arf1. We incubated the cells at 39.5 °C for 12 h to allow the VSV-G protein to accumulate in the ER, followed by a 20 min incubation at either 39.5 °C (**a**) or 32 °C (**b–g**). **h**, The abilities of either the purified recombinant Cdc42F28L mutant (GST–

Cdc42F28L) loaded with GTP γ S, Myc-tagged Cdc42F28L, or their corresponding diserine mutants (lysines 183 and 184 changed to serines) to bind endogenous coatomers assessed by western blotting with the anti α / γ COP antibody. The GST–Cdc42 proteins were incubated with NIH 3T3 cell lysates while the experiments with Myc-tagged Cdc42 were performed in COS-7 cells.

of its other known targets. The Cdc42Q61L diserine mutant (Cdc42Q61Lss) binds IQGAP as effectively as the activated Cdc42Q61L protein, and also binds putative targets that contain CRIB motifs including the PAKs, ACKs and WASP (Fig. 2c). The stimulation of PAK3 activity by Cdc42Q61Lss, as assayed by the phosphorylation of myelin basic protein, is indistinguishable from that of Cdc42Q61L (not shown). The Cdc42Q61Lss mutant also binds those regulators that associate with the activated form of the GTP-binding protein, namely the Cdc42-GTPase-activating protein (GAP) (Fig. 2c) and the Rho GDP-dissociation inhibitor (RhoGDI) (not shown). Thus, the γ COP subunit is the only identified Cdc42-binding partner that shows an altered affinity for the diserine mutant.

The involvement of coatomer assembly in intracellular trafficking^{10,11} raised the possibility that the specific binding of Cdc42 to the γ COP subunit might influence intracellular transport. The transmembrane glycoprotein of vesicular stomatitis virus (VSV-G) is a valuable model for studying intracellular transport because of the availability of temperature-sensitive mutants, which, at the permissive temperature, will transit from the endoplasmic reticulum (ER) through the Golgi on to the plasma membrane^{18,19}. Experiments using this system show that ER to Golgi transport is blocked by both nucleotide-exchange-defective and GTPase-defective mutants of Arf (ref. 20). This can be visualized when COS-7 cells expressing a temperature-sensitive mutant of VSV-G are transiently transfected with a cDNA encoding the GTPase-defective Arf1Q71L mutant. In control cells, VSV-G undergoes a

temperature-sensitive movement from the ER (Fig. 3a) to a more compact perinuclear Golgi location (Fig. 3e). However, in at least 75% of the cells that transiently expressed Arf1Q71L, the temperature-dependent transport of VSV-G was blocked, causing it to be retained in the ER (Fig. 3b). The transient expression of wild-type Cdc42 in COS-7 cells did not significantly inhibit transport (Fig. 3f). However, both the GTPase-defective Cdc42Q61L mutant, and the nucleotide-exchange-defective mutant Cdc42T17N blocked the transport of VSV-G out of the ER and into the Golgi as effectively as Arf1Q71L (Fig. 3c and d, respectively). We found the Golgi to be intact in cells expressing these Cdc42 mutants (not shown), indicating that this inhibition was not because of a disruption of the Golgi structure.

Given that the cycling of Cdc42 between the GDP- and GTP-bound states appeared to be necessary for normal trafficking activity, we examined the effects of a constitutively active Cdc42 mutant that was fully GTPase-competent (Cdc42F28L) on VSV-G transport activity. GST-Cdc42F28L, loaded with GTP γ S, formed a stable complex with the coatomer subunits *in vitro*. In addition, Myc-tagged Cdc42F28L, by virtue of its increased capability for existing in the GTP-bound state, co-immunoprecipitated with endogenous γ COP from COS-7 cells (Fig. 3h). In each case, the interaction was dependent on lysines 183 and 184 of Cdc42. As expected, the Cdc42F28L mutant did not block trafficking of VSV-G from the ER to Golgi (Fig. 3g); instead, we found that it accelerated transport activity. This is best illustrated by comparing the abilities of NIH 3T3 cells that stably expressed the Cdc42F28L mutant and

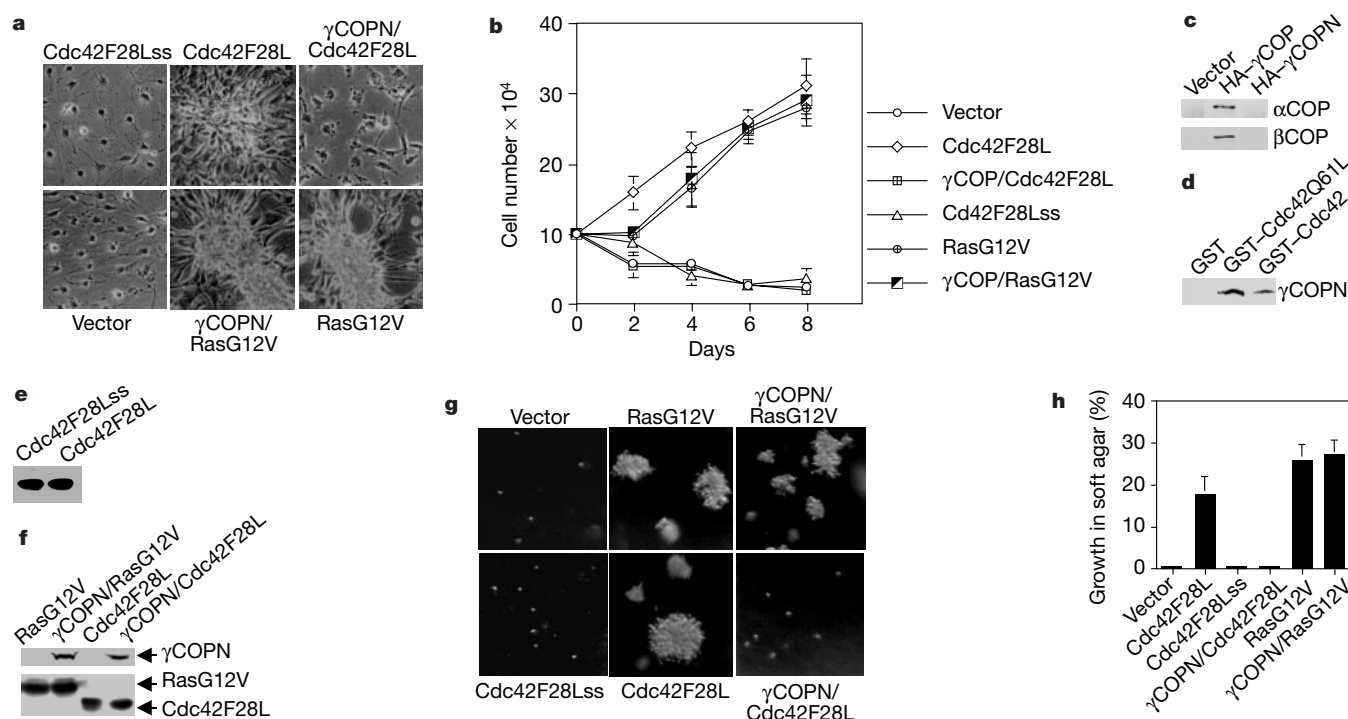


Figure 5 The ability of activated Cdc42 to bind to γ COP is essential for cellular transformation. **a**, Photomicrograph of cells stably expressing Cdc42F28Lss, Cdc42F28L, Cdc42F28L with the amino terminal half of γ COP (residues 1–554; designated γ COPN), control vector, oncogenic GTPase-defective Ras with γ COPN, and oncogenic Ras alone. We cultured the cells for 6 days in 1% serum. Original magnification is $\times 100$. **b**, Growth profiles (1% serum) for cells expressing the indicated forms of Cdc42 and Ras. Data represent the average of three experiments. **c**, HA-tagged γ COP expressed in COS-7 cells co-immunoprecipitates with endogenous α COP (western blotted with the α / γ COP antibody) and β COP (western blotted with anti- β COP antibody M3A5), but HA-tagged γ COPN does not. **d**, Recombinant activated GST-Cdc42Q61L, and to a lesser extent wild-type Cdc42, bind haemagglutinin (HA)-tagged γ COPN, assayed by precipitation with glutathione-agarose and western blotting with an anti-HA antibody. **e**, **f**, Relative

expression of Cdc42, Ras and γ COPN in NIH 3T3 cells. **e**, Western blot comparing relative expression of Cdc42F28L and Cdc42F28Lss. **f**, Western blot comparing the relative expression of RasG12V and Cdc42F28L in the presence and absence of γ COPN. **g**, Anchorage-independent growth of NIH 3T3 cells stably expressing vector alone, oncogenic Ras, oncogenic Ras plus γ COPN, Cdc42F28Lss, Cdc42F28L, and Cdc42F28L plus γ COPN. Photomicrograph of the colony formation of cells in soft agar. Original magnification is $\times 50$. **h**, Cell colonies were scored after 14 days for the conditions shown in **g**. Values shown are the average of three experiments. 100% represents 500 cells counted. In experiments comparing the effects of γ COPN on transformation, ten different colonies of Cdc42F28L/ γ COPN-expressing cells were examined and four different colonies of RasG12V/ γ COPN-expressing cells were assayed.

the Cdc42F28L diserine mutant to support VSV-G transport, as measured by the time-course for the proper processing (glycosylation) of the VSV-G protein (Fig. 4). The rate of VSV-G transport in cells expressing Cdc42F28L is significantly accelerated relative to cells stably expressing the diserine mutant. Cells expressing Cdc42F28L also consistently showed an increased rate of transport (1.5–2-fold), compared with control fibroblasts that expressed endogenous (wild-type) Cdc42 (Fig. 4, inset). We measured almost identical rates of VSV-G transport in cells expressing Cdc42 Δ L8F28L, a transformation-defective mutant that lacks the Rho-insert region²¹, and cells expressing the transforming Cdc42F28L mutant (Fig. 4). Thus, the stimulation of VSV-G transport by Cdc42F28L is not an indirect consequence of cell proliferation and transformation.

The results presented in Figs 3 and 4 imply that the activation of endogenous Cdc42 and its interaction with γ COP are necessary for some aspect of VSV-G transport. This may be related to the finding that activation of Cdc42 in MDCK cells is essential for the delivery of VSV-G protein to basolateral plasma membranes²². Consistent with these results, Cdc42 is predominantly found in the Golgi and its localization to the Golgi is affected by different Arf mutants in the same manner as the coatamer subunits²³. Perhaps, upon activation, Cdc42 competes with cargo receptors for binding to the γ COP subunit. GTP hydrolysis by Cdc42, together with Arf GTPase activity, could then facilitate the release of the coatamer subunits from transport vesicles and accelerate VSV-G trafficking. GTPase-defective Cdc42, by not releasing γ COP, would block uncoating and inhibit transport activity.

The requirement for Cdc42 to cycle between the GTP- and GDP-bound states to achieve maximal VSV-G transport activity and to stimulate cell growth and transformation⁴ raises an intriguing question. Is the Cdc42/ γ COP interaction essential for Cdc42-induced malignant transformation? NIH 3T3 cells that stably express Cdc42F28L exhibit a number of hallmarks of cellular transformation. These include the ability of Cdc42F28L-expressing cells to grow in 1% calf serum (Fig. 5a, compare bottom left with top middle; and Fig. 5b), and to form colonies in soft agar (Fig. 5g, compare top left with bottom middle; and Fig. 5h), an extremely reliable indicator of malignant transformation. The injection of Cdc42F28L-expressing cells into nude mice generates tumours with almost the same frequency and size as oncogenic Dbl (ref. 24). In contrast, cells expressing the Cdc42F28Ls diserine mutant did not grow in low serum (Fig. 5a, top left, and Fig. 5b) and were incapable of forming colonies in soft agar (Fig. 5g, bottom left, and Fig. 5h).

These findings lead to the conclusion that the ability of Cdc42 to bind to the γ COP subunit is necessary for Cdc42-induced malignant transformation. Given the indications that actin cytoskeletal organization, cell shape and locomotion are linked to intracellular transport^{25,26}, the hyper-activation of a Cdc42 trafficking function could lead to the enhancement of a cytoskeletal or growth regulatory signal that is essential for the transformed phenotype. The Cdc42- γ COP interaction is not sufficient for malignant transformation, as the Cdc42 Δ L8/F28L mutant, which is competent to bind γ COP and stimulate VSV-G transport (Fig. 4), is transformation-defective. As has been demonstrated for oncogenic Ras²⁷, the ability of activated Cdc42 to induce malignant transformation apparently requires the engagement of multiple targets. However, the binding of activated Cdc42 to γ COP causes a selective signal that distinguishes Ras- and Cdc42-mediated cellular transformation, as the C-terminally truncated γ COP molecule (γ COPN) is capable of binding activated Cdc42 (Fig. 5d) but is unable to bind the other coatamer subunits (Fig. 5c). The expression of γ COPN in cells that stably expressed Cdc42F28L (Fig. 5f) resulted in a marked block in Cdc42-mediated transformation (compare Cdc42F28L with γ COPN/Cdc42F28L in Fig. 5a, b, g and h). However, it had absolutely no effect on Ras-induced transformation (compare γ COPN/RasG12V with RasG12V in Fig. 5a, b, g and h). These

results provide further verification that Cdc42/ γ COP interactions are essential for Cdc42-stimulated transforming activity, as well as contributing to our understanding of the specific mechanisms by which Cdc42 and Ras transform cells²⁸. The distinct array of targets used by these two GTP-binding proteins indicates that Cdc42 and Ras use different signals to stimulate cell growth and promote cellular transformation. The identification of γ COP as a new target for Cdc42 highlights a previously unappreciated connection between the actions of Cdc42 in the Golgi and its ability to regulate cell growth and now provides a potentially important clue to the specificity underlying Cdc42 transforming activity. □

Methods

Identification of new binding partners for Cdc42

We incubated lysates from NIH 3T3 cells (usually 5 ml containing ~2 mg ml⁻¹ total protein) with glutathione-agarose beads containing GTPase-defective GST-Cdc42Q61L (~50 μ g), GST-Cdc42 (wild type) or GST alone, for 3 h at 4 °C with rocking. We pelleted the beads by centrifugation and resolved the proteins by SDS-polyacrylamide gel electrophoresis. We stained the gel with 1% Coomassie brilliant blue and excised the protein band (*M_r* 160K) that associated selectively with GTP-loaded GST-Cdc42Q61L, and analysed it by microsequencing at the Harvard Microchemistry Facility (Cambridge, Massachusetts).

VSV-G transport assays

We grew vesicular stomatitis virus (VSV) ts045 (a gift from A. Musch and E. Rodriguez-Boulan) in NIH 3T3 cells. We selected cell lines stably expressing Cdc42F28L, Cdc42 Δ L8/F28L and Cdc42F28Ls using G418 selection³ and infected them with 50 pfu per cell of VSV ts045 for 1 h at 37 °C. We incubated the cells for 3 h in DMEM plus 0.5% calf serum at 39.5 °C to accumulate VSV glycoprotein (VSV-G) in the endoplasmic reticulum and then washed them with methionine-free media (ICN Pharmaceuticals). We pulse-labelled the cells with [³⁵S]methionine (ICN Pharmaceuticals) for 30 min at 39.5 °C, followed by a chase incubation at 32 °C in the presence of 5 mM methionine and 20 μ g ml⁻¹ cycloheximide. We monitored the processing of labelled VSV-G as previously described¹⁹.

Indirect immunofluorescence

We plated COS-7 cells on the dual-chamber microscope slides (Nalge Nunc) in DMEM supplemented with 10% calf serum for 24 h, and then co-transfected them with 0.25 μ g of a plasmid encoding the temperature-sensitive mutant of VSV-G (pCDM8.1-VSV-G, a gift from J. Lippincott-Schwartz) and 0.5 μ g of the indicated HA-tagged Cdc42 constructs, using LipofectAMINE Plus TM reagent (GIBCO BRL). We incubated the transfected cells at 39.5 °C for 12 h followed by incubation at 39.5 °C or 32 °C for 20 min. We fixed the cells with 3.7% formaldehyde plus 0.2% Triton-X 100 in PBS. We used the mouse monoclonal anti-VSV-G antibody (P5D4) and the rabbit polyclonal anti-HA antibody (Covance) to detect VSV-G and HA-tagged Cdc42 proteins.

Studies with γ COPN

We constructed the dominant negative γ COP (γ COPN) by the deleting the C terminus of γ COP (amino acids 555–874) using a unique *Xma*I site. We co-transfected pJ4H γ COPN with a plasmid encoding a puromycin-resistance gene at a ratio of 20:1 into Cdc42F28L- or RasG12V-transformed NIH 3T3 cells. We selected cells with 5 μ g ml⁻¹ puromycin and screened puromycin-resistant colonies for the expression of HA-tagged Cdc42F28L or HA-tagged RasG12V and HA-tagged γ COPN.

Received 21 January; accepted 24 March 2000.

- Drubin, D. G. & Nelson, W. J. Origins of cell polarity. *Cell* **84**, 335–344 (1996).
- Johnson, D. I. Cdc42: An essential Rho-type GTPase controlling eukaryotic cell polarity. *Microbiol. Mol. Biol. Rev.* **63**, 54–105 (1999).
- Lin, R., Bagrodia, S., Cerione, R. A. & Manor, D. A novel Cdc42Hs mutant induces cellular transformation. *Curr. Biol.* **7**, 794–797 (1997).
- Hart, M. J., Eva, A., Evans, T., Aaronson, S. A. & Cerione, R. A. Catalysis of guanine nucleotide exchange on the CDC42Hs protein by the *dbf* oncogene product. *Nature* **354**, 311–314 (1991).
- Manser, E., Leung, T., Salihuddin, H., Zhao, Z.-S. & Lim, L. A brain serine/threonine protein kinase activated by Cdc42 and Rac1. *Nature* **367**, 40–46 (1994).
- Bagrodia, S., Taylor, S. J., Creasy, C. L., Chernoff, J. & Cerione, R. A. Identification of a mouse p21Cdc42/Rac activated kinase. *J. Biol. Chem.* **270**, 22731–22737 (1995).
- Symons, M. et al. Wiskott-Aldrich syndrome protein, a novel effector for the GTPase CDC42Hs, is implicated in actin polymerization. *Cell* **84**, 723–734 (1996).
- Hart, M. J., Callow, M. G., Souza, B. & Polakis, P. IQGAP1, a calmodulin-binding protein with a rasGAP-related domain, is a potential effector for cdc42Hs. *EMBO J.* **15**, 2997–3005 (1996).
- Serafini, T. et al. ADP-ribosylation factor is a subunit of the coat of Golgi-derived COP-coated vesicles: a novel role for a GTP-binding protein. *Cell* **67**, 239–253 (1991).
- Waters, M. G., Serafini, T. & Rothman, J. E. 'Coatamer': a cytosolic protein complex containing subunits of non-clathrin-coated Golgi transport vesicles. *Nature* **349**, 248–251 (1991).
- Rothman, J. E. Mechanisms of intracellular protein transport. *Nature* **372**, 55–63 (1994).
- Harter, C. et al. Nonclathrin coat protein gamma, a subunit of coatamer, binds to the cytoplasmic dylin motif of membrane proteins of the early secretory pathway. *Proc. Natl Acad. Sci. USA* **93**, 1902–1906 (1996).
- Lowe, M. & Kreis, T. E. *In vitro* assembly and disassembly of coatamer. *J. Biol. Chem.* **270**, 31364–31371 (1995).

14. Letourner, F. *et al.* Coatamer is essential for retrieval of dilysine-tagged proteins to the endoplasmic reticulum. *Cell* **79**, 1199–1207 (1994).
15. Harter, C. & Wieland, F. T. A single binding site for dilysine retrieval motifs and p23 within the gamma subunit of coatamer. *Proc. Natl Acad. Sci. USA* **95**, 11649–11654 (1998).
16. Bremser, M. *et al.* Coupling of coat assembly and vesicle budding to packaging of putative cargo receptors. *Cell* **96**, 495–506 (1999).
17. Erickson, J. W., Cerione, R. A. & Hart, M. J. Identification of an actin cytoskeletal complex that includes IQGAP and the Cdc42 GTPase. *J. Biol. Chem.* **272**, 24443–24447 (1997).
18. Kreis, T. Microinjected antibodies against the cytoplasmic domain of vesicular stomatitis virus glycoprotein block its transport to the cell surface. *EMBO J.* **5**, 931–941 (1986).
19. Musch, A., Xu, H., Shields, D. & Rodriguez-Boulton, E. Transport of vesicular stomatitis virus G protein to the cell surface is signal mediated in polarized and nonpolarized cells. *J. Cell Biol.* **133**, 543–558 (1996).
20. Dascher, C. & Balch, W. E. Dominant inhibitory mutants of ARF1 block endoplasmic reticulum to Golgi transport and trigger disassembly of the Golgi apparatus. *J. Biol. Chem.* **269**, 1437–1448 (1994).
21. Wu, W., Lin, R., Cerione, R. A. & Manor, D. Transformation activity of Cdc42 requires a region unique to Rho-related proteins. *J. Biol. Chem.* **273**, 16655–16658 (1998).
22. Kroschewski, R., Hall, A. & Mellman, I. Cdc42 controls secretory and endocytic transport to the basolateral plasma membrane of MDCK cells. *Nature Cell Biol.* **1**, 8–13 (1999).
23. Erickson, J. W., Zhang, C. J., Kahn, R. A., Evans, T. & Cerione, R. A. Mammalian Cdc42 is a brefeldin A-sensitive component of the Golgi apparatus. *J. Biol. Chem.* **271**, 26850–26854 (1996).
24. Lin, R., Cerione, R. A. & Manor, D. Specific contributions of the small GTPases Rho, Rac, and Cdc42 to Dbp transformation. *J. Biol. Chem.* **274**, 23633–23641 (1999).
25. Bershadsky, A. D. & Futerma, A. H. Disruption of the Golgi apparatus by brefeldin A blocks cell polarization and inhibits directed cell migration. *Proc. Natl Acad. Sci. USA* **91**, 5686–5689 (1994).
26. Peranen, J., Auvinen, P., Virta, H., Wepf, R. & Simons, K. Rab8 promotes polarized membrane transport through reorganization of actin and microtubules in fibroblasts. *J. Cell Biol.* **135**, 153–167 (1996).
27. White, M. A. *et al.* Multiple Ras functions can contribute to mammalian cell transformation. *Cell* **80**, 533–541 (1995).
28. Westwick, J. K. *et al.* Rac regulation of transformation, gene expression, and actin organization by multiple, PAK-independent pathways. *Mol. Cell Biol.* **17**, 1324–1335 (1997).
29. McCallum, S. J., Erickson, J. W. & Cerione, R. A. Characterization of the association of the actin-binding protein, IQGAP, and activated Cdc42 with Golgi membranes. *J. Biol. Chem.* **273**, 22537–22544 (1998).
30. Bagrodia, S., Taylor, S. J., Jordan, K. A., Van Aelst, L. & Cerione, R. A. A novel regulator of p21-activated kinases. *J. Biol. Chem.* **273**, 23633–23636 (1998).

Acknowledgements

We thank C. Harter and F. Wieland for the anti- α/γ COP antibody and the cDNA for γ COP; A. Musch for advice on VSV-G transport assays; J. Lippinott-Schwartz for the plasmid expressing VSV-G; and C. Westmiller for secretarial assistance. We also acknowledge support from the National Institutes of Health.

Correspondence and requests for materials should be addressed to R.A.C. (e-mail: rac1@cornell.edu).

Two-headed binding of a processive myosin to F-actin

Matthew L. Walker*, Stan A. Burgess*, James R. Sellers†, Fei Wang‡, John A. Hammer III‡, John Trinick* & Peter J. Knight*

* Astbury Centre for Structural Molecular Biology, School of Biomedical Sciences, University of Leeds, Leeds LS2 9JT, UK

† Laboratories of Molecular Cardiology and ‡ Cell Biology, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

Myosins are motor proteins in cells. They move along actin by changing shape after making stereospecific interactions with the actin subunits¹. As these are arranged helically, a succession of steps will follow a helical path. However, if the myosin heads are long enough to span the actin helical repeat (~36 nm), linear motion is possible. Muscle myosin (myosin II) heads are about 16 nm long², which is insufficient to span the repeat³. Myosin V, however, has heads of about 31 nm that could span 36 nm (refs 4, 5) and thus allow single two-headed molecules to transport cargo by walking straight⁵. Here we use electron microscopy to show that while working, myosin V spans the helical repeat. The heads are mostly 13 actin subunits apart, with values of 11 or 15 also found.

Typically the structure is polar and one head is curved, the other straighter. Single particle processing reveals the polarity of the underlying actin filament, showing that the curved head is the leading one. The shape of the leading head may correspond to the beginning of the working stroke of the motor. We also observe molecules attached by one head in this conformation.

Myosin V is a class of motor protein implicated in vesicle transport in cells and is particularly abundant in neurons¹. The kinetics of its ATPase show differences from those of myosin II (muscle myosin) which may be adaptations for hand-over-hand 'processive' action^{6,7}. Thus the rate of ADP release from the complex with actin is close to the overall cycle rate, whereas ATP binding and cleavage are much faster. The affinity of myosin V–ADP for actin is very high. These properties ensure that the heads are attached to actin for a larger fraction of the ATPase cycle than myosin II, that is, their duty ratio is higher. For fully processive motion by this mechanism, one head should not detach before the second re-attaches. Although this is the case for two-headed motors that move along microtubules^{8,9}, it is controversial whether myosin V typically takes many or few steps before detaching^{10,11}.

The structural basis of the processive movement has not been established. Early electron microscopy of the complex of myosin V and actin found bundles of filaments crosslinked by the two heads⁴, in the manner previously seen with myosin II (ref. 12). When excess myosin V was present, fully decorated actin was found with a fringe of material about 50 nm wide. Some of this was fine strands, probably the tails, but some appeared more globular, suggestive of molecules attaching by a single head. Binding of both heads to a single actin filament was not reported. The two heads of myosin II can attach to the same filament, and are generally on adjacent subunits, only occasionally on the next but one³. To understand processive movement, it is therefore important to show whether, and under what conditions, the two heads of myosin V can bind to a single actin filament.

Optical trap records show that at 1 μ M ATP, steps occur at about one-second intervals¹⁰. This rate is limited by the rate of ATP binding, and we reasoned that this should favour observation of molecules attached by both heads: the trailing one being free of nucleotide, the lead one probably at an earlier stage in the cycle. To see the structure of individual molecules attached to actin we mixed a low molar ratio of myosin V HMM (heavy meromyosin, that is, lacking the cargo-binding domain) with actin under these conditions. This myosin has a high affinity for actin, so we were able to avoid ATP depletion by using only 40 nM HMM; we also used ~100 mM ionic strength, in contrast to the requirements for observing myosin II during ATPase activity¹³. We observed about 50% of the bound molecules attached by both heads (Fig. 1a–d). The heads are spaced far apart along actin, as required for linear processive function, and attachment to adjacent actin subunits was not observed. There were also many singly attached (Fig. 1e) and unattached molecules, and many of the actin filaments were bundled together. Myosin V purified from chick brain¹⁴ and myosin V HMM stored in excess calmodulin gave very similar results (data not shown), indicating that the wide separation of the attached heads is not an artefact of a misfolded or calmodulin-depleted preparation. Doubly-attached heads became rare if the grid was rinsed with a higher concentration of ATP, as expected if the trailing head is free of nucleotide.

In the absence of ATP or ADP, we found a strong tendency towards bundling of actin filaments by myosin V HMM, and the structure within these bundles was obscure. By using more dilute proteins and allowing only a few seconds between mixing and adsorption to the carbon film, we reduced bundling. We observed very few unattached molecules under these conditions, indicating that almost all the myosin was competent in actin binding. Many myosin molecules were attached by only one head, but the doubly-attached molecules appeared similar to those in Fig. 1a–d. The

presence of molecules with only one head attached, and the strong bundling tendency, indicate that it is relatively difficult for two nucleotide-free myosin V heads to bind to the same actin filament. In rigor the two heads have the same preferred conformation and there would have to be substantial distortion in the second head to allow binding to the same filament.

In the presence of saturating ATP (1.0 mM), we observed many more unattached molecules, indicating that myosin V HMM is not strongly processive. Those that were attached bound almost exclusively by one head, and bundling was infrequent.

ADP slows the rate of actomyosin V ATPase by forming an actin–myosin–ADP intermediate. As this is a tight binding complex for myosin V (ref. 6), ADP should favour two-headed binding. In the absence of ATP, 20 μ M ADP reduced bundling compared to rigor, and more doubly-attached molecules were present. This indicates that the actin binding properties of myosin V HMM–ADP are different from those of nucleotide-free heads. Addition of 2 μ M ADP to 1 μ M ATP caused an increase in bundling, but had little

effect on the incidence of double attachment. Addition of 100 μ M ADP to 100 μ M ATP increased the fraction of molecules attached, the amount of bundling and the incidence of doubly-attached molecules. Thus the incidence of double attachment follows the expectations from the kinetic properties of myosin V (refs 6, 7).

To determine the separation between the heads of doubly-attached molecules we applied single particle image processing^{15–17}, using the actin subunit periodicity to calibrate the magnification of the micrographs. Figure 2 shows that the heads were typically 13 actin subunits apart, a distance of 36 nm. With 13/6 actin helical symmetry, these strides will produce linear motion. Figure 2 also shows subsidiary peaks of 11 and 15 subunits between attached heads (30 nm and 41 nm, respectively). These subunits are azimuthally adjacent to subunit 13. Molecules with different stride lengths appeared similar in other respects (Fig. 1b–d). This variation may be a consequence of cumulative angular disorder in actin filaments¹⁸, sometimes bringing subunits 11 or 15, rather than subunit 13, to the same azimuth as the first. Flexibility within the myosin may also contribute, allowing heads to vary the azimuth as well as the distance at which they attach. Variation in stride length may thus contribute to the variation in step size measured in optical trap experiments.

Most doubly-attached molecules (70%) show an obvious polarity because both heads are curved in the same direction, though to different degrees, like a skier's legs in the telemark stance (Fig. 1a–d). Where more than one telemark-shaped molecule was attached to a single actin filament they all showed the same polarity (Fig. 1a), indicating that this structure is not merely a random distortion. To determine the polarity of the telemark shape relative to that of F-actin, we aligned short (48 nm) segments of F-actin to which heads were attached to generate a polar average (Fig. 1f). Inspection of the HMM molecules attached to the aligned segments showed that for the great majority (82% of 203 molecules) the barbed end of the filament was beyond the strongly curved head. As the direction of movement of this

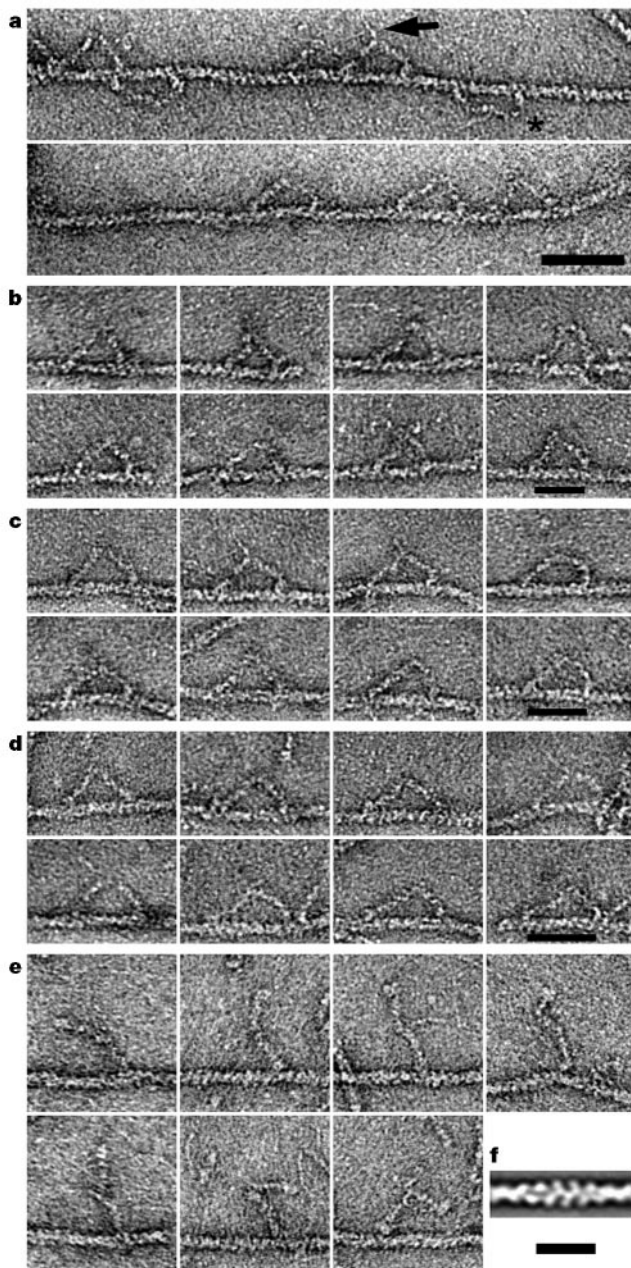


Figure 1 Structure of myosin V HMM bound to F-actin. All filaments are shown with the barbed end to the right. **a**, Filaments with several examples of two-headed binding in 1 μ M ATP. Arrow, an HMM tail pointing away from the barbed end of the actin; asterisk, W-shaped molecule. **b–d**, Galleries of single HMM molecules spanning 11, 13 and 15 actin subunits respectively; left column shows symmetrical structure, the others the polar telemark structure, orientated with the leading head to the right; bars correspond to a distance of 11 (**b**), 13 (**c**) or 15 (**d**) subunits. **e**, Examples of single-headed binding in 1 μ M ATP, arranged to simulate a plausible walking cycle (see Fig. 4; also animated at <http://www.leeds.ac.uk/bms/research/muscle/muscle.htm>). **f**, Typical average of aligned segments of F-actin underlying myosin V heads. Scale bar in **a** 50 nm; applies to **a–e**, **f**, 20 nm. We expressed the HMM-like fragment of mouse myosin Va heavy chain⁷ (residues 1–1,091) and calmodulin by co-infection of Sf9 cells by baculovirus vectors, purified them as described⁷ and stored them at $\sim 1.6 \mu$ M in 0.5 M KCl, 0.1 mM EGTA, 1 mM dithiothreitol, 10 mM MOPS, pH 7.0. Actin was from rabbit skeletal muscle, stored in liquid N₂ at 220 μ M in 0.2 mM CaCl₂, 0.5 mM DTT, 0.2 mM ATP, 2 mM Tris–HCl, pH 8.0 at 0 °C. We formed the Mg polymer by addition of 0.2 mM EGTA and 0.1 mM MgCl₂ followed after 15 min by 50 mM KCl. We diluted the proteins using 0.10 M KCl, 1.0 mM EGTA, 1.0 mM MgCl₂, 10 mM MOPS, pH 7.0. We made specimens from equal volumes of F-actin (0.4–1.6 μ M) and myosin V HMM (0.08–0.16 μ M), mixed and immediately applied to carbon-film EM grids as described²⁵. We stained the grids with 1% uranyl acetate, generally without rinsing as this washed away cycling molecules. For rigor and ADP experiments, we treated 40 μ M F-actin with 0.01 units ml^{–1} apyrase (Sigma A6535) for 5 min before use. When needed, we incubated 40 μ M ADP for ~ 2 min with the HMM only before mixing with actin. When used, we added ATP to both proteins 5 s before they were mixed. We calibrated micrographs at $\times 40,000$ using Fourier transforms of straight segments of F-actin, and scanned (Leafscan-45) at a step size corresponding to 0.53 nm at the specimen. Image processing was performed using the SPIDER software suite²⁶.

myosin along actin is towards the barbed end⁴, this indicates that the strongly curved head leads and the straighter one trails. We suppose that the remainder, with reversed polarity, originate from incorrectly assigned actin polarity, as these were more numerous where the actin structure correlated weakly with the average. With the direction of movement determined, we can also observe that the short HMM tail is typically angled in the trailing direction (for example, Fig. 1a, arrow). About 30% of doubly-attached molecules show sufficiently little difference in shape between the heads that polarity is not apparent (Fig. 1b–d, left). Molecules with this structure were found in all experiments. Molecules are also seen that are W- rather than V-shaped (Fig. 1a, asterisk). We suspect that these are attached to actin at azimuths that place them out of the plane of the carbon film, and that their appearance is a consequence of collapse.

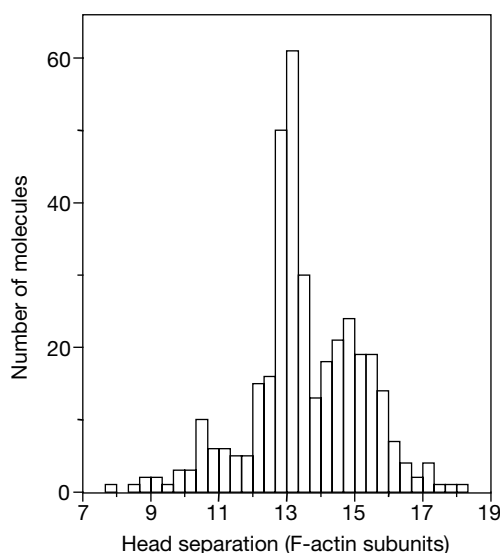


Figure 2 Axial distance spanned by two heads of a myosin V HMM molecule attached to actin in 1 μ M ATP. The distance between heads was measured by aligning the motor domains and using the resulting rotation and translation parameters to determine their positions in the original micrographs. $n = 367$. Specimens prepared in 20 μ M ADP also gave a peak separation of 13 subunits.

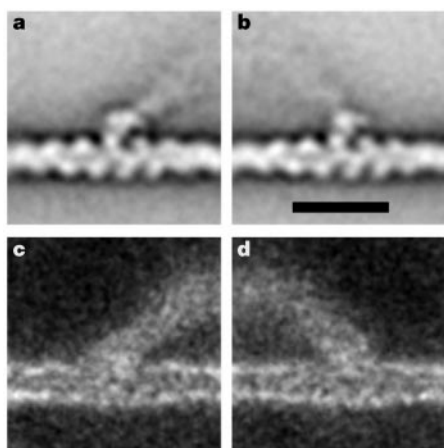


Figure 3 Average and variance images respectively of 262 lead (b, d) and 234 trail (a, c) heads of myosin V HMM bound to actin in the presence of 1 μ M ATP. Scale bar: 20 nm. We aligned the underlying actin first, then aligned the lead and trail heads using a mask that essentially excluded the actin. The clarity of the actin structure indicates that these averages derive from heads bound to a narrow range of actin azimuths.

Image averaging (Fig. 3) shows that both motor domains are primarily associated with single actin subunits and are similar in shape. The actin subunits to which they attach appear the same. Variation in shape between molecules weakens the light chain domains in the averages, but the variance images (Fig. 3c, d) indicate the range of shapes. The trailing light chain domain emerges from the leading side of the motor domain and is angled at roughly 40° to the filament (Fig. 3a); this resembles the rigor angle¹⁹, thought to be the end of the working stroke. The averaged structure of the leading head is quite different as its light chain domain emerges near the trailing side of the motor domain and is strongly angled back to join the other head. The angle to the filament, measured near the junction of the light chain domain with the motor domain, varies between individual images with a range of 90–150°; from the average image (Fig. 3b), the angle is about 115°. Comparing the averages there is thus a pronounced (75°) difference between the two heads in the orientation of the light chain domain emerging from the motor domain, and this is in the direction expected to produce a forward working stroke by the leading head once the trailing head detaches. Rotation through this angle would produce an axial translation of 26 nm for the head–tail junction of an unconstrained head. Laser trap measurements of single working strokes by myosin V give similar values (~20 nm)^{11,20}, which indicates that the images are relevant to the start and end of the working stroke.

The telemark-shaped molecules may show us two key states of the processive motion (Fig. 4). Evidence is accumulating for a strain dependence in ADP release in this and other myosins^{6,19,21}. The structure we have observed for the leading head is that predicted to correspond to the direction of strain that would inhibit ADP release, whereas the trailing head is strained in the direction likely to promote release. Thus it is plausible that in images obtained at low ATP concentrations, the trailing head is nucleotide-free and the leading head may still contain ADP. Detachment of the trailing head by ATP allows the leading head to tilt forward, carrying the detached

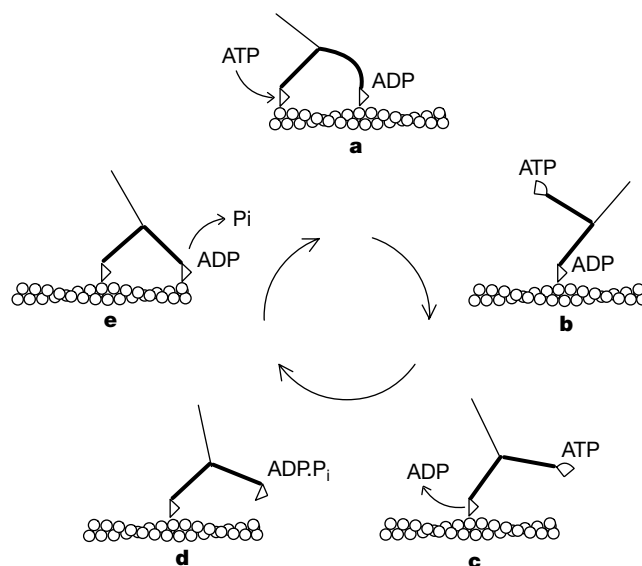


Figure 4 Model showing the role of two-headed attachment in the walking cycle of myosin V. The trailing, nucleotide-free head of a telemark-shaped molecule binds ATP (a) and detaches, relieving the axial strain in the molecule (b) and bringing the detached head towards the next actin site (c). Dissociation of ADP completes the working stroke (d). Attachment of the free head is quickly followed by loss of phosphate by actin activation, and a tight binding develops (e). The detached head is shown in two conformations to indicate the transitions to a weak binding, followed by a primed conformation indicated by crystal structures of myosin II heads^{23,24}.

head towards its next binding site. Rebinding could be facilitated by the sharp angle induced between motor and lever arm on ATP binding^{22–24}, as this would resemble the structure we see for the leading head. The shape of the leading head is the start of the walking stride, and the trailing head the end.

The relationship between the working stroke of a single myosin V head and the 36-nm walking stride of the two-headed molecule requires clarification. As mentioned, the working stroke has been estimated at ~20 nm in the laser trap^{11,20} and ~26 nm here. These values fall well short of the ~36 nm span seen here, indicating that some additional flexibility in the heads allows the detached head to diffuse forward and attach further along the filament than the working stroke has carried it. A stride of 36 nm correlates well with the average steps (~34–38 nm) measured in optical trap experiments when a succession of steps is seen¹⁰. This indicates that the structural basis of stepping is a succession of strides between sites on actin rather than the working strokes of the heads themselves. In molecules attached by a single head, a variety of attachment angles is apparent (Fig. 1e). After assignment of the actin polarity, some are near the rigor angle, but others are found to be tilted in the same direction as the lead head of the doubly-attached molecules, and therefore appear to be molecules caught at the start of their working stroke.

Note added in proof: The doubly attached conformation seen here is similar to a recently published atomic model of this complex extrapolated from crystal structure data of actin subunits and myosin II heads in rigor and ADP–AlF₄ states²⁷. □

Received 10 February; accepted 19 April 2000.

1. Sellers, J. R. *Myosins* (Oxford Univ. Press, Oxford, 1999).
2. Rayment, I. *et al.* Three-dimensional structure of myosin subfragment-1: a molecular motor. *Science* **261**, 50–58 (1993).
3. Craig, R. *et al.* Electron microscopy of thin filaments decorated with a Ca²⁺-regulated myosin. *J. Mol. Biol.* **140**, 35–55 (1980).
4. Cheney, R. E. *et al.* Brain myosin-V is a two-headed unconventional myosin with motor activity. *Cell* **75**, 13–23 (1993).
5. Howard, J. Molecular motors: structural adaptations to cellular functions. *Nature* **389**, 561–567 (1997).
6. De La Cruz, E. M., Wells, A. L., Rosenfeld, S. S., Ostap, E. M. & Sweeney, H. L. The kinetic mechanism of myosin V. *Proc. Natl Acad. Sci. USA* **96**, 13726–13731 (1999).
7. Wang, F. *et al.* Effect of ADP and ionic strength on the kinetic and motile properties of recombinant mouse myosin V. *J. Biol. Chem.* **275**, 4329–4335 (2000).
8. Howard, J., Hudspeth, A. J. & Vale, R. D. Movement of microtubules by single kinesin molecules. *Nature* **342**, 154–158 (1989).
9. Rice, S. *et al.* A structural change in the kinesin motor protein that drives motility. *Nature* **402**, 778–784 (1999).
10. Mehta, A. D. *et al.* Myosin-V is a processive actin-based motor. *Nature* **400**, 590–593 (1999).
11. Veigel, C. *et al.* Is one myosin V molecule sufficient for vesicular transport? *Biophys. J.* **78**, 242A (2000).
12. Trinick, J. & Offer, G. Cross-linking of actin filaments by heavy meromyosin. *J. Mol. Biol.* **133**, 549–556 (1979).
13. Walker, M., Trinick, J. & White, H. Millisecond time resolution electron cryo-microscopy of the M-ATP transient kinetic state of the acto-myosin ATPase. *Biophys. J.* **68**, 87S–91S (1995).
14. Cheney, R. E. Purification and assay of myosin V. *Methods Enzymol.* **298**, 3–18 (1998).
15. Frank, J. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies* (Academic, New York, 1996).
16. Burgess, S. A., Walker, M. L., White, H. D. & Trinick, J. Flexibility within myosin heads revealed by negative stain and single-particle analysis. *J. Cell Biol.* **139**, 675–681 (1997).
17. Burgess, S. A. *et al.* Real-space 3-D reconstruction of frozen-hydrated arthrin and actin filaments at 2 nm resolution. *Biophys. J.* **78**, 8A (2000).
18. Egelman, E. H. & DeRosier, D. J. Image analysis shows that variations in actin crossover spacings are random, not compensatory. *Biophys. J.* **63**, 1299–1305 (1992).
19. Whittaker, M. *et al.* A 35-A movement of smooth muscle myosin on ADP release. *Nature* **378**, 748–751 (1995).
20. Moore, J. R., Kremontsova, E., Trybus, K. M. & Warshaw, D. M. Myosin V exhibits a high duty cycle and large unitary displacement at zero load. *Biophys. J.* **78**, 272A (2000).
21. Veigel, C. *et al.* The motor protein myosin-I produces its working stroke in two steps. *Nature* **398**, 530–533 (1999).
22. Sugimoto, Y., Tokunaga, M., Takezawa, Y., Ikebe, M. & Wakabayashi, K. Conformational changes of the myosin heads during hydrolysis of ATP as analyzed by x-ray solution scattering. *Biophys. J.* **68**, 29S–34S (1995).
23. Smith, C. A. & Rayment, I. X-ray structure of the magnesium(II). ADP. vanadate complex of the Dictyostelium discoideum myosin motor domain to 1.9 Å resolution. *Biochemistry* **35**, 5404–5417 (1996).
24. Houdusse, A., Kalabokis, V. N., Himmel, D., Szent-Györgyi, A. G. & Cohen, C. Atomic structure of scallop myosin subfragment S1 complexed with MgADP: a novel conformation of the myosin head. *Cell* **97**, 459–470 (1999).
25. Walker, M., Knight, P. & Trinick, J. Negative staining of myosin molecules. *J. Mol. Biol.* **184**, 535–542 (1985).

26. Frank, J., Shimkin, B. & Dowse, H. SPIDER- a modular software system for electron image-processing. *Ultramicroscopy* **6**, 343–357 (1981).
27. Vale, R. D. & Milligan, R. A. The way things move: looking under the hood of molecular motor proteins. *Science* **288**, 88–95 (2000).

Acknowledgements

We thank C. Veigel, H. White and S. Schmitz for discussions and E. Harvey for technical assistance supported by BBSRC (to J.T. and P.J.K.) and NIH (to J.T. and H. White).

Correspondence and requests for materials should be addressed to P.J.K. (e-mail: p.j.knight@leeds.ac.uk). An animated sequence of images is available at <http://www.leeds.ac.uk/bms/research/muscle/muscle.htm>.

The lyase activity of the DNA repair protein β -polymerase protects from DNA-damage-induced cytotoxicity

Robert W. Sobol, Rajendra Prasad, Andrea Evenski, Audrey Baker, Xiao-Ping Yang, Julie K. Horton & Samuel H. Wilson

Laboratory of Structural Biology, National Institute of Environmental Health Sciences, PO Box 12233, Research Triangle Park, North Carolina 27709-2233, USA

Small DNA lesions such as oxidized or alkylated bases are repaired by the base excision repair (BER) pathway¹. BER includes removal of the damaged base by a lesion-specific DNA glycosylase, strand scission by apurinic/apyrimidinic endonuclease, DNA resynthesis and ligation². BER may be further subdivided into DNA β -polymerase (β -pol)-dependent single-nucleotide repair and β -pol-dependent or -independent long patch repair subpathways^{3–6}. Two important enzymatic steps in mammalian single-nucleotide BER are contributed by β -pol: DNA resynthesis of the repair patch and lyase removal of 5'-deoxyribose phosphate (dRP)². Fibroblasts from β -pol null mice are hypersensitive to monofunctional DNA-methylating agents, resulting in increases in chromosomal damage, apoptosis and necrotic cell death^{3,7}. Here we show that only the dRP lyase activity of β -pol is required to reverse methylating agent hypersensitivity in β -pol null cells. These results indicate that removal of the dRP group is a pivotal step in BER *in vivo*. Persistence of the dRP moiety in DNA results in the hypersensitivity phenotype of β -pol null cells and may signal downstream events such as apoptosis and necrotic cell death.

Several DNA polymerases can support single-nucleotide and long patch BER DNA synthesis *in vitro*^{4,5}, indicating that other endogenous polymerases might functionally complement the β -pol deficiency in β -pol null cells. However, β -pol null cells are hypersensitive to the cytotoxic effects of the methylating agent methyl methanesulphonate (MMS)³, indicating an absolute *in vivo* requirement for β -pol. Here we investigate this essential role of β -pol in resistance to methylating agents.

We constructed β -pol minigenes³ to express Flag epitope-tagged variants of the enzyme that have wild-type activity or are deficient in polymerase activity, lyase activity or both (Fig. 1c). These minigenes were introduced into β -pol null cells and the resulting MMS hypersensitivity phenotype was analysed. Cells expressing wild-type β -pol (β -pol null/WT β -pol cells) were as resistant to MMS as wild-type cells, whereas control cells expressing the neomycin resistance gene (β -pol null/Neo cells) and untransfected cells were hypersensitive to MMS (Fig. 1a and b). The roles of the carboxy-terminal polymerase domain (relative molecular mass 31,000; M_r

31K) and the amino-terminal DRP lyase domain (M_r 8K)² in resistance to MMS were investigated using the β -pol variants shown in Fig. 1c. These variants express the 8K and 31K domains of β -pol, respectively (not shown), or mutants with single or multiple enzymatic active-site point mutations. Expression of β -pol from these minigenes was at least as high as in the wild-type cell line MB36.3 (not shown).

The DNA synthesis active site of β -pol has been characterized and the residues involved in nucleotidyltransferase activity are known. In particular, Arg 283 is critical for efficient nucleotide incorporation and discrimination^{8,9}. In the β -pol/substrate crystal structures, Arg 283 forms a hydrogen bond and makes van der Waals' contact with the template strand in the DNA minor groove, and is proposed to stabilize the template base during recognition of the incoming dNTP^{8,10}. Mutation of this residue to alanine results in a polymerase with a 5,000-fold decreased catalytic efficiency in DNA synthesis. We expressed this mutant polymerase (R283A) in β -pol null cells. As shown in Fig. 1d (lane 2), the immunoaffinity-purified Flag-R283A- β -pol protein was deficient in BER DNA synthesis, as expected⁸, but had full DRP lyase activity (Fig. 1e, lane 2).

Unexpectedly, β -pol null/R283A- β -pol cells showed wild-type sensitivity to MMS (Fig. 2a, filled circles), whereas the control β -pol/Neo cells (Fig. 2a, open circles) behaved like β -pol null cells. There was no apparent requirement for the DNA synthesis activity of β -pol to complement the β -pol null phenotype. To examine further the requirement for β -pol DNA synthesis activity, we generated a cell line expressing a polymerase active-site deletion

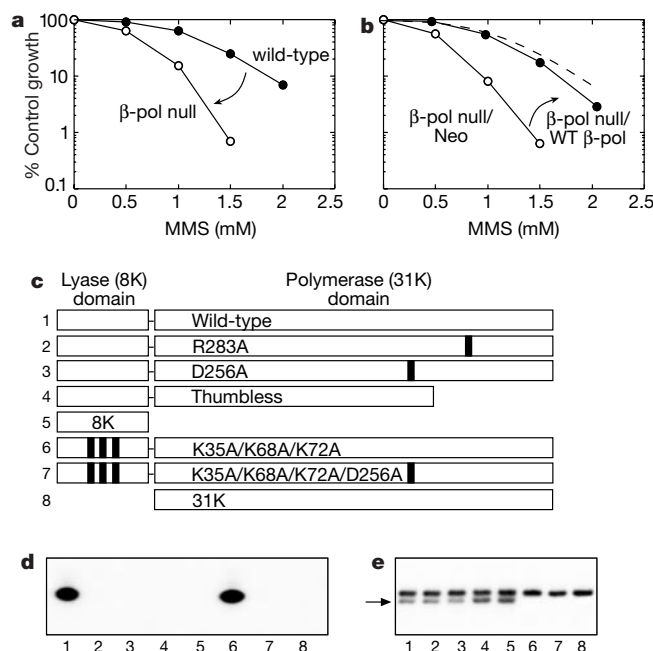


Figure 1 Sensitivity of fibroblast cell lines to MMS. Survival curves in the presence of MMS of wild-type cells (filled circles) and β -pol null cells (open circles) (**a**), and β -pol null/WT β -pol cells (filled circles) and β -pol null/Neo cells (open circles) (**b**). Arrow, shift of the curve due to the genetic loss (**a**) or gain (**b**) of β -pol. Dashed line, cytotoxic sensitivity curve for wild-type cells when directly compared in the same experiment. **c**, Diagram of recombinant β -pol proteins expressed in β -pol null cells. Vertical bars, approximate locations of point mutations. Numbers on the left indicate the corresponding lane in **d** and **e**. **d**, DNA synthesis activity determined for Flag-wild-type β -pol (lane 1) and Flag- β -pol mutants (lanes 2–8) using immunoaffinity purified proteins in a single-nucleotide gap-filling assay (see Methods). **e**, DRP lyase activity determined using immunoaffinity purified Flag-wild-type β -pol (lane 1) and Flag- β -pol mutant proteins (lanes 2–8) using a pre-cisec dsDNA oligonucleotide (see Methods). Arrow, reaction product. In **d** and **e**, lane 1, Flag-wild-type β -pol; lane 2, Flag-R283A β -pol; lane 3, Flag-D256A β -pol; lane 4, Flag-Thumbless; lane 5, Flag-8K; lane 6, Flag-K35A,K68A,K72A β -pol; lane 7, Flag-K35A,K68A,K72A,D256A β -pol; lane 8, Flag-31K.

mutant of β -pol (Flag-Thumbless), in which residues within the C-terminal portion (residues 263–335) of β -pol that are required for DNA synthesis activity are deleted^{8–11}. As Asp 256 is critical to the nucleotidyltransferase mechanism, and substitution of alanine for aspartate (D256A) abolishes the polymerase activity¹², we also generated a cell line expressing a polymerase active-site point mutant, Flag-D256A- β -pol. Both Flag-D256A- β -pol and Flag-Thumbless proteins were completely deficient in single-nucleotide BER gap-filling DNA synthesis (Fig. 1d, lanes 3 and 4), but were active in DRP removal (Fig. 1e, lanes 3 and 4). We tested β -pol null cells expressing either the Thumbless mutant or the D256A mutant (Fig. 1c) for MMS sensitivity. Both β -pol null/Thumbless β -pol cells (Fig. 2b, closed circles) and β -pol null/D256A- β -pol cells (not shown) showed a sensitivity to MMS similar to that of wild-type cells. These observations indicate that DNA synthesis catalysed by β -pol is not critical for cell survival following MMS treatment. This result is similar to the earlier observation that the catalytic activity of yeast DNA polymerase- ϵ is dispensable for replication and repair^{13,14}, but the protein itself is essential. Thus, the data presented here indicate that the β -pol DNA synthesis activity is not essential for resistance to MMS, whereas the DRP lyase activity is.

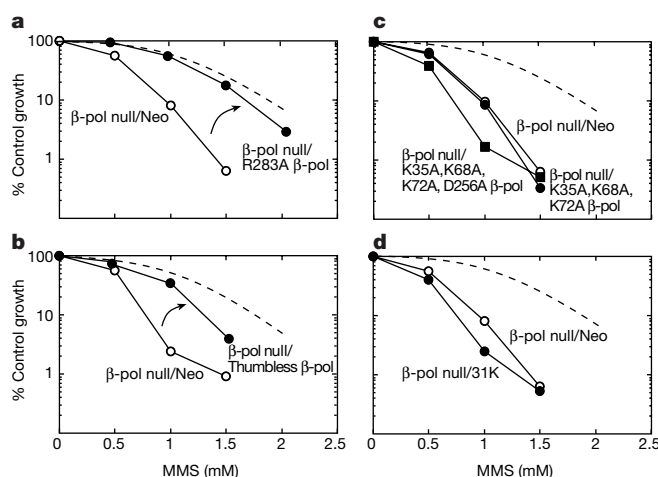


Figure 2 DNA synthesis activity of β -pol is not required to reverse the MMS hypersensitivity of β -pol null cells. **a**, MMS sensitivity of β -pol null/R283A β -pol cells (filled circles) and β -pol null/Neo cells (open circles). **b**, MMS sensitivity of β -pol null/Thumbless β -pol cells (filled circles) and β -pol null/Neo cells (open circles). **c**, MMS sensitivity of β -pol null/K35A,K68A,K72A β -pol cells (filled circles), β -pol null/K35A,K68A,K72A,D256A β -pol cells (filled squares) and β -pol null/Neo cells (open circles). **d**, MMS sensitivity of β -pol null/31K β -pol cells (filled circles) and β -pol null/Neo cells (open circles). Dashed line, cytotoxic sensitivity curve for wild-type cells when directly compared in the same experiment.

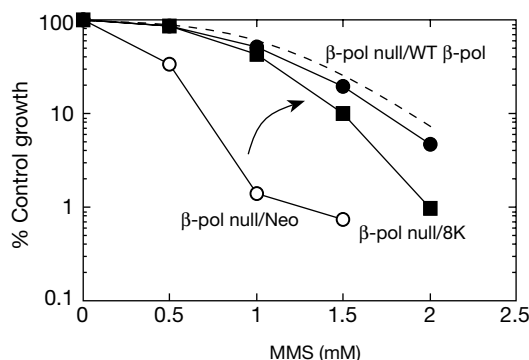


Figure 3 DRP lyase activity of β -pol is sufficient to reverse the MMS hypersensitivity of β -pol null cells. MMS sensitivity of β -pol null/WT β -pol cells (filled circles), β -pol null/8K β -pol cells (filled squares) and β -pol null/Neo cells (open circles). Dashed line, cytotoxic sensitivity curve for wild-type cells when directly compared in the same experiment.

To determine whether dRP lyase activity is critical for resistance to MMS, β -pol null cells were stably transfected with a full-length β -pol minigene with three alanine point mutations within the dRP lyase active site. Lys 72 is the predominant Schiff base nucleophile in the dRP lyase β -elimination reaction mechanism, and mutating three active site lysine residues to alanine (K35A, K68A, K72A) completely eliminates dRP lyase activity^{15–17}. Flag-K35A,K68A,K72A- β -pol was expressed in β -pol null cells and immunoaffinity purified to characterize its DNA synthesis and dRP lyase activities *in vitro*. Flag-K35A,K68A,K72A- β -pol was proficient in DNA synthesis (Fig. 1d, lane 6), but it had no dRP lyase activity (Fig. 1e, lane 6). When β -pol null/K35A,K68A,K72A- β -pol cells were exposed to increasing doses of MMS, there was no change in MMS hypersensitivity when compared with that of β -pol null/Neo cells (Fig. 2c, filled circles). In addition, the Flag-K35A,K68A,K72A,D256A β -pol mutant, which is deficient in both dRP lyase and DNA synthesis activities (Fig. 1d and e), was expressed in β -pol null cells; these cells were also similar to β -pol null/Neo cells in their sensitivity to MMS (Fig. 2c, filled squares).

The β -pol protein can be separated into the N-terminal 8K dRP lyase and the C-terminal 31K DNA polymerase domains. The C-terminal 31K domain (Flag-31K, Fig. 1c) has little DNA binding activity and diminished DNA synthesis activity². Purified Flag-31K is deficient in gap-filling DNA synthesis and dRP lyase activity (Fig. 1d and e, respectively). As with the other dRP lyase mutants described above, β -pol null/31K β -pol cells were no more resistant to MMS than β -pol null or β -pol null/Neo cells (Fig. 2d).

The N-terminal 8K dRP lyase domain of β -pol has single-

stranded DNA binding activity, 5'-phosphate recognition and dRP lyase enzymatic activity². Flag-8K was expressed in β -pol null cells and the protein was immunoaffinity purified. This small β -pol domain had no gap-filling DNA synthesis activity when measured *in vitro* (Fig. 1d), but, as expected, exhibited dRP lyase activity (Fig. 1e)^{15,16}. As the entire C-terminal domain has been removed, the Flag-8K protein cannot have DNA synthesis activity (Fig. 1c). To determine whether the 8K domain is sufficient to restore MMS resistance, β -pol null/Neo, β -pol null/WT β -pol and β -pol null/8K β -pol cells were treated with MMS. As shown in Fig. 3, β -pol null/8K β -pol cells had wild-type resistance to MMS (filled squares). Note the apparent increase in sensitivity of the β -pol null/8K β -pol cells at higher MMS concentrations as compared with the β -pol null/WT β -pol cells. This indicates that at high concentrations of MMS direct dRP removal by the 8K domain of β -pol may not be entirely sufficient for cell survival. It is possible that the alternative long patch BER pathway is more significant at high doses of MMS, requiring the DNA synthesis activity of β -pol¹⁸.

The observed correlation between failure to repair the 5'-dRP group and hypersensitivity to MMS was unexpected, because other proteins with 5'-dRP lyase activity have been observed in mammalian cells^{19,20}. In addition, dRP lyase activity has been found in various *Escherichia coli*, *Drosophila* and yeast proteins, many of which have mammalian homologues^{21–25}. The potential contribution of these other dRP lyase activities towards the 5'-dRP-containing BER intermediate was assessed. Whole-cell extracts were prepared from wild-type and β -pol null cells, β -pol null/WT β -pol cells, β -pol null cells expressing the β -pol lyase-proficient point mutants Flag-R283A and Flag-D256A, β -pol null/8K β -pol cells and β -pol null cells expressing the lyase-deficient mutant Flag-K35A,K68A,K72A β -pol. We measured the capacity of these extracts to remove the 5'-dRP lesion from the BER intermediate in a complete BER assay using a double-stranded DNA oligonucleotide substrate with a single G-U base pair. As outlined in Fig. 4a, base removal and cleavage 5' to the abasic site is conducted by endogenous uracil DNA glycosylase and AP endonuclease, respectively, giving rise to the 5'-dRP-containing BER intermediate. Extracts from all the cell lines expressing dRP lyase-proficient β -pol proteins (wild-type, Flag-WT β -pol, Flag-R283A β -pol and Flag-D256A β -pol; Fig. 4b, lanes 1–4) had a high capacity to remove the dRP group from the AP endonuclease-incised substrate (upper band), thereby producing the 29-mer (lower band) product. On the other hand, cells deficient in β -pol (Fig. 4b, lanes 5 and 6) or cells expressing a β -pol dRP lyase-deficient mutant (Flag-K35A,K68A,K72A β -pol; Fig. 4b, lane 7) did not exhibit dRP lyase activity. Finally, an extract from the β -pol null/8K β -pol cells showed high BER-associated dRP lyase activity (Fig. 4b, lane 8).

Our results give rise to the following hypothesis: the 5'-dRP group in the BER intermediate is a cytotoxic lesion produced after MMS exposure, and cellular proteins other than β -pol do not efficiently repair this 5'-dRP group. The removal of dRP is the rate-limiting step during single-nucleotide BER²⁶, and the 8K domain of β -pol is responsible for this essential function. However, another cellular polymerase can substitute for β -pol to incorporate a single nucleotide residue in the repair of MMS-induced DNA damage. It remains to be determined whether repair of DNA lesions induced by other DNA-damaging agents is also absolutely dependent on the dRP lyase activity of β -pol. □

Methods

Plasmids

We subcloned the full-length Flag-WT β -pol minigene³ into pBluescript II SK- (Stratagene). Single point mutations were generated by PCR with primers containing the desired point mutations. We generated multiple point mutations by multiple rounds of amplification and sequenced all cloned PCR products. Each mutant mini-gene was cloned into (1) pRS1021, a mammalian expression plasmid with a single lox P site 5' to the CMV promoter derived from pcDNA3 (Invitrogen), (2) pIRENeo (Clontech) or (3) pIRE-Spuro (Clontech).

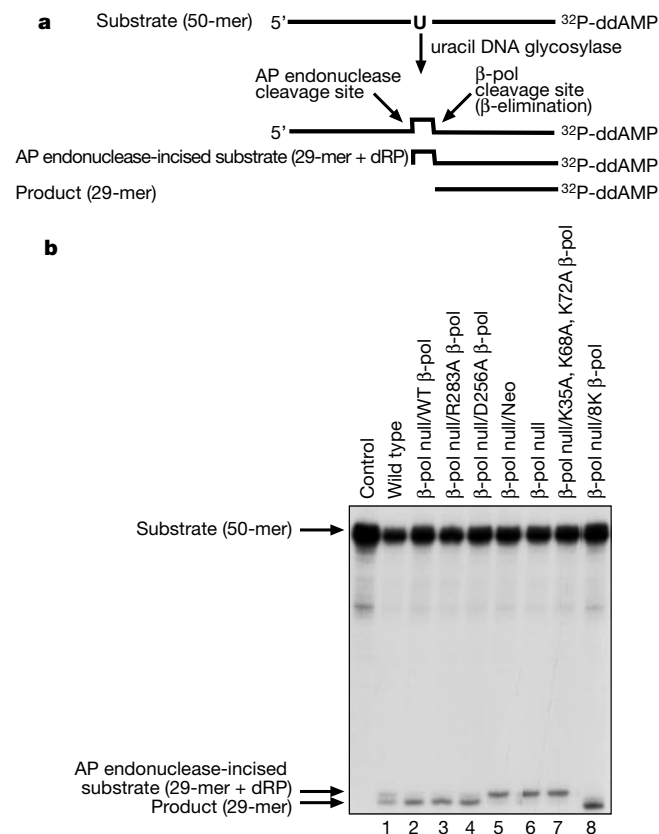


Figure 4 dRP lyase activity in whole-cell extracts from stably transfected β -pol null cell lines. Extracts were from wild-type mouse fibroblasts (lane 1), β -pol null fibroblasts (lanes 5 and 6) or β -pol null fibroblasts expressing wild-type or mutant Flag- β -pol proteins as indicated (lanes 2–4, 7, 8). **a**, Assay was carried out with a ³²P-labelled (3') 50-bp dsDNA substrate containing a single G-U base pair (see Methods). Only one DNA strand is shown. **b**, A representative autoradiograph. The position of the substrate, the AP endonuclease-incised substrate and the product are indicated.

Cell culture

MB36.3 (wild-type) and MB38Δ4 cells (β-pol null) are SV40 T-Ag transformed cell lines that harbour the λ phage transgene *λLIZ*²⁷. We transfected cells³ with the indicated plasmids and selected stable transformants with G418 (600 μg ml⁻¹) or puromycin (6 μg ml⁻¹). Clones carrying pIRESneo or pIRESpuo were screened by western blotting and clones carrying pRS1021 integrated at the β-pol locus were screened by PCR.

Immunoaffinity purification and western blotting

We prepared immunoaffinity purified proteins from extracts of transfected cells. Cells (10⁶ cells per 20 μl) were suspended in buffer I (10 mM Tris-Cl, pH 7.8; 200 mM KCl) and an equal volume of buffer II was added (10 mM Tris-Cl, pH 7.8, 200 mM KCl, 2 mM EDTA, 40% glycerol, 0.2% Nonidet P-40, 2 mM dithiothreitol, 0.5 mM phenylmethylsulphonyl fluoride, 10 mg ml⁻¹ aprotinin, 5 mg ml⁻¹ leupeptin, 1 mg ml⁻¹ pepstatin). The suspension was rotated at 4 °C for 1 h and centrifuged at 14,000 r.p.m. for 10 min. The supernatant was recovered and 20 μg anti-Flag M2 antibody conjugated to agarose (Kodak) was added to the cell extract, incubated with rotation for 6 h at 4 °C and pelleted by centrifugation at 14,000 r.p.m. (15 s). The supernatant was discarded and the protein-bound agarose complex was washed three times in extract buffer. The Flag-β-pol proteins were eluted by addition of Flag-peptide at 150 μg ml⁻¹ and incubation for 30 min at 4 °C. The eluted Flag-β-pol was stored at -80 °C. Total cellular protein (20 μg) was separated by electrophoresis in a 4–12% gradient SDS-polyacrylamide gel (Novex) and electrotransferred to a nitrocellulose membrane. We detected β-pol by incubating the membrane with rabbit anti β-pol polyclonal antibody followed by horseradish peroxidase (HRP)-conjugated secondary antibody (Santa Cruz). HRP activity was detected by enhanced chemiluminescence (Pierce).

Gap-filling assay

Immunoaffinity-purified proteins were assayed using a single-nucleotide gap-filling assay. The indicated proteins were incubated in buffer containing 50 mM HEPES, pH 7.8, 2 mM EDTA, 1 mM dithiothreitol, 10 mM MgCl₂, 100 mM KCl, 330 nM [α³²P]-dCTP and 100 nM single-nucleotide gap substrate, constructed by annealing three oligodeoxynucleotides (5' upper strand, 5'-CTGCAGCTGATGCGC-3'; 3' upper strand, 5'-pGTACGGATCCCCGGGTAC-3'; lower strand, 5'-GTACCGGGGATCCGTACGGCGC ATCAGCTGCGAC-3', to yield a dsDNA substrate with a single nucleotide gap at position 21 and a 5'-phosphate in the gap. The reaction was incubated for 2 min at 37 °C and analysed by electrophoresis in a 16% polyacrylamide gel (7 M urea, TBE). Gels were fixed, dried and autoradiographed.

dRP lyase activity

Immunoaffinity-purified proteins were assayed for dRP lyase activity using a ³²P-labelled (3') 50-base-pair (bp) DNA substrate containing a single abasic site pre-incised with AP endonuclease as described^{15,16}. Whole-cell extracts were assayed with a ³²P-labelled (3') 50-bp DNA substrate containing a single G:U base pair, as outlined in Fig. 4a. Before the assay, whole-cell extracts were dialysed against reaction buffer (50 mM HEPES, pH 7.5, 0.1 mM EDTA, 100 mM KCl, 1 mM dithiothreitol, 20% glycerol). In both assays, the reaction products were stabilized by addition of NaBH₄ and then separated by electrophoresis in a 20% polyacrylamide gel (7 M urea, TBE). Gels were fixed, dried and autoradiographed.

Cytotoxicity assay

Cytotoxicity was determined by growth inhibition assays as described^{3,6}. Results are the mean of at least two separate experiments. % Control growth is (number of treated cells)/(number of control cells) × 100.

Received 7 February; accepted 18 April 2000.

- Lindahl, T. & Wood, R. D. Quality control by DNA repair. *Science* **286**, 1897–1905 (1999).
- Wilson, S. H. Mammalian base excision repair and DNA polymerase β. *Mutat. Res.* **407**, 203–215 (1998).
- Sobol, R. W. *et al.* Requirement of mammalian DNA polymerase β in base-excision repair. *Nature* **379**, 183–186 (1996).
- Biade, S., Sobol, R. W., Wilson, S. H. & Matsumoto, Y. Impairment of proliferating cell nuclear antigen-dependent apurinic/aprimidinic site repair on linear DNA. *J. Biol. Chem.* **273**, 898–902 (1998).
- Fortini, P. *et al.* Different DNA polymerases are involved in the short- and long-patch base excision repair in mammalian cells. *Biochemistry* **37**, 3575–3580 (1998).
- Horton, J. K., Prasad, R., Hou, E. & Wilson, S. H. Protection against methylation-induced cytotoxicity by DNA polymerase β-dependent long patch base excision repair. *J. Biol. Chem.* **275**, 2211–2218 (2000).
- Ochs, K., Sobol, R. W., Wilson, S. H. & Kaina, B. Cells deficient in DNA polymerase β are hypersensitive to alkylating agent-induced apoptosis and chromosomal breakage. *Cancer Res.* **59**, 1544–1551 (1999).
- Beard, W. A. *et al.* Enzyme–DNA interactions required for efficient nucleotide incorporation and discrimination in human DNA polymerase β. *J. Biol. Chem.* **271**, 12141–12144 (1996).
- Ahn, J., Werneburg, B. G. & Tsai, M. D. DNA polymerase β: structure-fidelity relationship from pre-steady-state kinetic analyses of all possible correct and incorrect base pairs for wild type and R283A mutant. *Biochemistry* **36**, 1100–1107 (1997).
- Beard, W. A. & Wilson, S. H. Structural insights into DNA polymerase β fidelity: hold tight if you want it right. *Chem. Biol.* **5**, 7–13 (1998).
- Pelletier, H., Sawaya, M. R., Kumar, A., Wilson, S. H. & Kraut, J. Structures of ternary complexes of rat DNA polymerase β, a DNA template-primer, and ddCTP. *Science* **264**, 1891–1903 (1994).
- Menge, K. L. *et al.* Structure–function analysis of the mammalian DNA polymerase β active site: role of aspartic acid 256, arginine 254, and arginine 258 in nucleotidyl transfer. *Biochemistry* **34**, 15934–15942 (1995).

- Kesti, T., Flick, K., Keranen, S., Syvaoja, J. E. & Wittenberg, C. DNA polymerase ε catalytic domains are dispensable for DNA replication, DNA repair, and cell viability. *Mol. Cell* **3**, 679–685 (1999).
- Dua, R., Levy, D. L. & Campbell, J. L. Analysis of the essential functions of the C-terminal protein/protein interaction domain of *Saccharomyces cerevisiae* pol ε and its unexpected ability to support growth in the absence of the DNA polymerase domain. *J. Biol. Chem.* **274**, 22283–22288 (1999).
- Prasad, R. *et al.* Functional analysis of the amino-terminal 8-kDa domain of DNA polymerase β as revealed by site-directed mutagenesis. DNA binding and 5'-deoxyribose phosphate lyase activities. *J. Biol. Chem.* **273**, 11121–11126 (1998).
- Prasad, R., Beard, W. A., Strauss, P. R. & Wilson, S. H. Human DNA polymerase β deoxyribose phosphate lyase. Substrate specificity and catalytic mechanism. *J. Biol. Chem.* **273**, 15263–15270 (1998).
- Matsumoto, Y., Kim, K., Katz, D. S. & Feng, J. A. Catalytic center of DNA polymerase β for excision of deoxyribose phosphate groups. *Biochemistry* **37**, 6456–6464 (1998).
- Prasad, R., Dianov, G. L., Bohr, V. A. & Wilson, S. H. FEN1 stimulation of DNA polymerase β mediates an excision step in mammalian long patch base excision repair. *J. Biol. Chem.* **275**, 4460–4466 (2000).
- Price, A. & Lindahl, T. Enzymatic release of 5'-terminal deoxyribose phosphate residues from damaged DNA in human cells. *Biochemistry* **30**, 8631–8637 (1991).
- Longley, M. J., Prasad, R., Srivastava, D. K., Wilson, S. H. & Copeland, W. C. Identification of 5'-deoxyribose phosphate lyase activity in human DNA polymerase γ and its role in mitochondrial base excision repair in vitro. *Proc. Natl Acad. Sci. USA* **95**, 12244–12248 (1998).
- Sandigursky, M. & Franklin, W. A. *Escherichia coli* single-stranded DNA binding protein stimulates the DNA deoxyribosephosphodiesterase activity of exonuclease I. *Nucleic Acids Res.* **22**, 247–250 (1994).
- Sandigursky, M. *et al.* The yeast 8-oxoguanine DNA glycosylase (Ogg1) contains a DNA deoxyribosephosphodiesterase (dRpase) activity. *Nucleic Acids Res.* **25**, 4557–4561 (1997).
- Sandigursky, M., Yacoub, A., Kelley, M. R., Deutsch, W. A. & Franklin, W. A. The *Drosophila* ribosomal protein S3 contains a DNA deoxyribosephosphodiesterase (dRpase) activity. *J. Biol. Chem.* **272**, 17480–17484 (1997).
- Dianov, G. *et al.* Release of 5'-terminal deoxyribose-phosphate residues from incised abasic sites in DNA by the *Escherichia coli* RecJ protein. *Nucleic Acids Res.* **22**, 993–998 (1994).
- Graves, R. J., Felzenszwalb, I., Laval, J. & O'Connor, T. R. Excision of 5'-terminal deoxyribose phosphate from damaged DNA is catalyzed by the Fpg protein of *Escherichia coli*. *J. Biol. Chem.* **267**, 14429–14435 (1992).
- Srivastava, D. K. *et al.* Mammalian abasic site base excision repair. Identification of the reaction sequence and rate-determining steps. *J. Biol. Chem.* **273**, 21203–21209 (1998).
- Dycaico, M. J. *et al.* The use of shuttle vectors for mutation analysis in transgenic mice and rats. *Mutat. Res.* **307**, 461–478 (1994).

Acknowledgements

We thank W. A. Beard and K. R. Tindall for critical reading of the manuscript.

Correspondence and requests for materials should be addressed to S.H.W. (e-mail: wilson5@niehs.nih.gov).

Observations of light-induced structural changes of retinal within rhodopsin

Gerhard Gröbner*, Ian J. Burnett, Clemens Glaubitz, Gregory Choi†, A. James Mason & Anthony Watts

Biomembrane Structure Unit, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK

Photo-isomerization of the 11-*cis* retinal chromophore activates the mammalian light-receptor rhodopsin¹, a representative member of a major superfamily of transmembrane G-protein-coupled receptor proteins (GPCRs) responsible for many cell signal communication pathways. Although low-resolution (5 Å) electron microscopy studies^{2,3} confirm a seven transmembrane helix bundle as a principal structural component of rhodopsin, the structure of the retinal within this helical bundle is not known in detail. Such information is essential for any theoretical or functional understanding of one of the fastest occurring photoactivation processes in nature, as well as the general mechanism behind GPCR activation^{4–6}. Here we determine the three-dimensional

* Present address: Department of Chemistry, Biophysical Chemistry, Umeå University, SE-901 87 Umeå, Sweden.

† Present address: University of Connecticut, Molecular & Cell Biology, LSA 205, 75 North Eagleville Road, U-125 Storrs, Connecticut 06269-3125, USA.

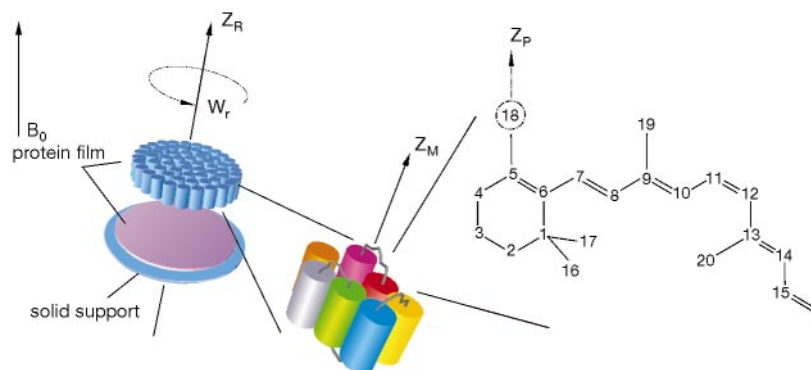


Figure 1 In the magic angle oriented sample spinning (MAOSS) NMR approach used here, uni-axially oriented phospholipid membranes containing bovine rhodopsin were stacked on glass plates in a MAS rotor, with rotor axis Z_R parallel to the sample director. The

chromophore retinal (shown as 11-*cis* retinal in its 6-*s-trans* form) carried a deuterated methyl group at position 18 whose orientation could be determined.

structure of 11-*cis* retinal bound to bovine rhodopsin in the ground state at atomic level using a new high-resolution solid-state NMR method⁷. Significant structural changes are observed in the retinal following activation by light to the photo-activated M_1 state of rhodopsin giving the all-*trans* isomer of the chromophore. These changes are linked directly to the activation of the receptor, providing an insight into the activation mechanism of this class of receptors at a molecular level.

In the mammalian photoreceptor rhodopsin, the external stimulus is an optical one and it controls the activation of a GTP-protein-coupled signal transduction pathway leading to a neural response¹. More usually, however, GPCRs transmit chemical (hormone, pharmacologically mediated) signals across cellular membranes^{4,8}. Even though GPCRs share a putative common spatial protein arrangement^{4,5}, no member of this family, other than rhodopsin^{2,3-6}, has yet been resolved in any significant ($<5 \text{ \AA}$) structural detail, which explains its importance in any structural and functional studies for the whole family of GPCRs to reveal their putative common mechanism of activation. The lack of high-resolution ($<5 \text{ \AA}$) structural information for rhodopsin means that no comprehensive structural description at the molecular level is available

for 11-*cis* retinal, the light sensitive prosthetic group of rhodopsin whose interaction with its receptor binding pocket causes rhodopsin activation. The conjugated polyene chain of retinal and its five methyl groups and cyclohexene ring (Fig. 1) all interact with various amino acids of the retinal binding pocket, and so may also induce conformational changes within the protein upon light-induced isomerization. Also, the seven-helix bundle arrangement of the protein undergoes a conformational change to activate the GTP-protein transducin⁹ followed by a cascade of signalling events. Numerous biophysical studies have provided a significant body of indirect and often inconsistent evidence for the retinal environment without resolving the molecular detail of either the retinal binding site or the retinal conformation itself, or the structural features of the molecular mechanism of light activation^{1,2,9-14}. More recently, solid state NMR studies have provided some direct and highly localized details of the Schiff base environment and the non-planarity in the polyene chain¹⁴⁻¹⁸. However, a complete picture of the structure and orientation of retinal at an atomic level has still not emerged, and information essential for any theoretical and functional understanding of the general mechanism behind rhodopsin activation is far from complete.

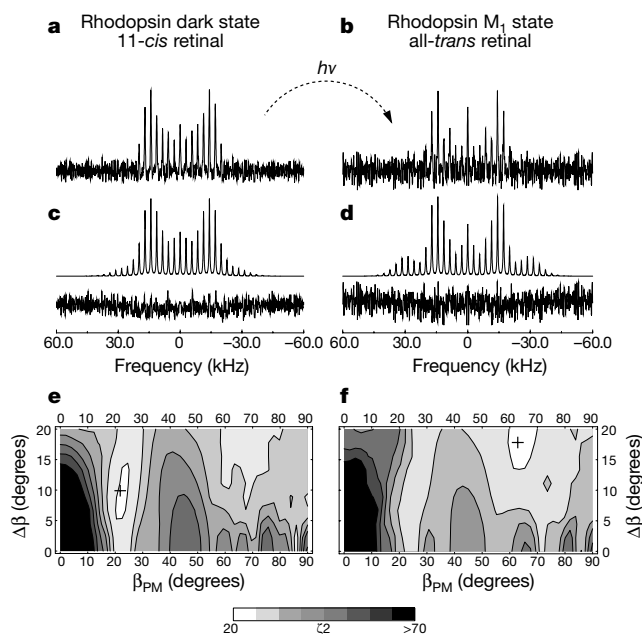


Figure 2 Deuterium-MAOSS NMR spectra of $[18\text{-C}(^2\text{H})_3]$ -retinal in dark adapted rhodopsin (a) and upon photo-activation in the M_1 state (illuminated below 273 K) (b), at $\omega_r = 2860 \text{ Hz}$ and $T = 213 \text{ K}$. The spectra were analysed by minimizing the r.m.s. deviation of the MAS sideband intensities in the $[\beta_{PM}, \Delta\beta]$ -parameter space, as shown by the χ^2 -

contour plots (e) and (f). The $\text{C}-\text{C}(^2\text{H})_3$ tilt angle is determined to be $\beta_{PM} = 21 \pm 5^\circ$ in the dark adapted state and $62 \pm 7^\circ$ in the M_1 state. It can be seen that $\Delta\beta$ changes slightly from 10° in the ground state to 18° in M_1 . The difference between best fit and experimental spectra are shown in (c) and (d).

To determine the three-dimensional structure and alignment of retinal in the binding pocket of rhodopsin at the atomic level, high-resolution solid state deuterium (^2H) NMR methods were employed, namely static^{19,20} and the newly developed magic angle oriented sample spinning (MAOSS) approach⁷. The suggested 11-*cis* retinal structure in the ground state of the receptor is distorted¹⁷, with the ring moiety twisted against the polyene chain and further twists around the $\text{C}_{10}=\text{C}_{11}$ double bond. Therefore, we synthesized three different retinals^{7,20}, which are all analogues of the naturally occurring chromophore but with each of the following methyl groups, C_5-C_{18} , C_9-C_{19} and $\text{C}_{13}-\text{C}_{20}$, carrying deuterium $\text{C}-\text{C}(^2\text{H})_3$, to be used separately as non-perturbing NMR reporters of molecular structure. The solid state NMR spectra recorded from each of the labelled retinals in rhodopsin contain direct information about the orientation of the labelled $\text{C}-\text{C}(^2\text{H})_3$ bond with respect to the applied magnetic field^{7,20} and directly reveal the orientational constraints for each of the various molecular segments of the chromophore within the protein and the changes induced after isomerization by light.

Representative ^2H MAOSS NMR spectra are shown in Fig. 2a and b for rhodopsin containing a $\text{C}_5-\text{C}_{18}(^2\text{H})_3$ group on the cyclohexene ring of the retinal, in the ground state and upon photo-activation in its thermally trapped M_1 state. Each sideband intensity is a function of size and orientation of the deuterium quadrupolar coupling tensor, allowing a precise, direct and reliable measurement of the $\text{C}-\text{C}(^2\text{H})_3$ bond vector orientation to be made, as illustrated with the root-mean-square deviation plots of the $\text{C}_5-\text{C}_{18}(^2\text{H})_3$ orientation with respect to the membrane normal. In this way a comprehensive analysis provides information not only about the orientation of the chromophore in the binding pocket of rhodopsin, but also about the intramolecular changes occurring upon photo-activation in the M_1 state. Significant changes in the bond vectors are observed upon photo-excitation of retinal within rhodopsin, in particular in the ring moiety where the C_5-C_{18} bond has an orientation of $21 \pm 5^\circ$ in the inactivated state with respect to the membrane normal, changing markedly to $62 \pm 7^\circ$ for the M_1 state (Table 1), this being the intermediate immediately before major protein conformational changes occur.

The detailed intramolecular structural information obtained for retinal within the protein is resolved without the need to rely on any knowledge of the protein structure itself. In particular, details can be resolved for the orientation of the ring moiety of retinal, and its twist relative to the conjugated polyene chain, a topic which has been the centre of much debate substantiated mainly by studies using various optical absorption and ^{13}C -NMR chemical shift data^{10,11,21}, even though a complete description of the chromophore was not available. Here, the MAOSS approach provides tight constraints (bond vectors for $\text{C}-\text{C}(^2\text{H})_3$ determined to $\pm 5-7^\circ$) on the orientation of the ring moiety in the protein body and its twist relative to the other individual molecular segments. From static ^2H NMR¹⁹, the orientational constraints for the C_{19} methyl group are $44 \pm 5^\circ$ and for the C_{20} group are $30 \pm 5^\circ$. Using these direct orientational constraints, and the torsional and distance constraints measured by solid-state NMR for the $\text{C}_{10}-\text{C}_{11}-\text{C}_{12}-\text{C}_{13}$ segment¹⁶⁻¹⁸, we have constructed the first complete high-resolution, three-dimensional structure of the orientation and conformation of 11-*cis* retinal in the binding pocket of rhodopsin (Fig. 3).

Major differences are observed for the position of the cyclohexene ring with respect to the polyene chain when compared with earlier

suggested structural models^{10,13,14,21}. To accommodate the tilt angle of $21 \pm 5^\circ$ obtained here for the C_5-C_{18} bond vector, a rotational movement of the ring must be performed around the C_6-C_7 bond axis, which itself is tilted by 50° from the protein long axis. This procedure results in two possible positions of the ring, with either a twist of 28° around the C_6-C_7 axis or of 70° (corresponding to a torsion angle of -28° and -70° for the $\text{C}_1-\text{C}_6-\text{C}_7-\text{C}_8$ segment). The latter (shown in Fig. 3) being an angle which is more likely to reduce steric interaction of the ring methyl groups attached to the C_1 atom with the C_8 proton because of an increase in separation of more than $0.6 \text{ \AA}$, thereby drastically reducing the violation of van der Waals distances. In both cases the retinal has a 6-*s-trans* as previously found for retinal in bacteriorhodopsin²² and not the 6-*s-cis* configuration assumed from indirect optical and ^{13}C NMR chemical shift studies^{10,11,13,21}.

In a theoretical calculation of chiro-optical properties for bound 11-*cis* retinal, the observed circular dichroism (CD) spectra show very little dependence on the rotation of the β -ionone ring with respect to conjugated polyene chain¹¹. The ^{13}C NMR chemical shift value for the C_5 ring atom reported for retinal in rhodopsin is found at the upper end of the region for 6-*s-cis* model compounds, close to the spectral range for 6-*s-trans* compounds. However, the chemical shift value is extremely sensitive to several factors including the degree of conjugation between the $\text{C}_5=\text{C}_6$ double bond and the conjugated bond system of the polyene chain, and the electronic properties of amino-acid residues surrounding the ring moiety. Furthermore, a twist angle of 70° for a 6-*s-trans* configuration of the retinal ring provides a similar degree of conjugational overlap as a ring-chain twist of $60-70^\circ$ for a 6-*s-cis* configuration. This situation is very different from the planar 6-*s-trans* retinal model compounds used as references in the NMR studies²¹. Thus, reconciling the structurally and electronically sensitive optical and NMR chemical shift data reported elsewhere^{10-14,21} with the direct and structurally sensitive deuterium NMR presented here, is very difficult.

The average orientation of the retinal axis (defined from C_6 to C_{12}) with respect to the membrane plane was found to be 22° from the current study, which is less than the 32° measured by indirect optical (and hence averaged) approaches¹⁰⁻¹³. The rotational angle of 55° for the retinal plane (defined as $\text{C}_6-\text{C}_9-\text{C}_{19}$) is slightly larger than the $40-50^\circ$ range indicated by optical studies¹³. However, the discrepancies may result simply from the fact that the retinal environment contains a complex electronic distribution that con-

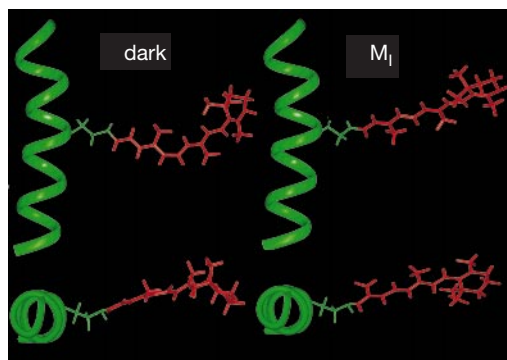


Figure 3 The structure and orientation of the chromophore 11-*cis* retinal in the binding pocket of the GPCR protein bovine rhodopsin in its dark-adapted ground state (left), including the experimentally obtained structural constraints (helix 7 of the protein used as z-axis reference; helix only partially displayed). The steric and electrostatic interactions with the protein induce an out of plane deformation of the chromophore and twists along the molecule. The structure and orientation of the chromophore all-*trans* retinal in the binding pocket of the GPCR protein bovine rhodopsin in its trapped M_1 state upon photo-activation is shown on the right. The top shows a view along membrane plane with cytoplasmic side of the disc membrane above, and the bottom is a view from cytoplasmic side of protein down the long axis of helix 7 on to the chromophore.

Table 1 Summary of NMR spectra analysis

	C_5-C_{18}	C_9-C_{19}	$\text{C}_{13}-\text{C}_{20}$
Dark state	$21 \pm 5^\circ$	$44 \pm 5^\circ$	$30 \pm 5^\circ$
M_1 state	$62 \pm 7^\circ$	$65 \pm 10^\circ$	$60 \pm 10^\circ$

Bond vectors for each of the labelled segments of retinals within bovine rhodopsin measured directly from the ^2H -NMR spectra.

tributes significantly to measured optical spectra, but does not give direct structural constraints for the individual molecular chromophoric segments separately. These details are, however, available from the electronic-insensitive but structurally sensitive deuterium NMR approach used here.

The position of the chromophore with respect to the membrane bilayer has been resolved (Fig. 3) using information from the helix wheel representation for rhodopsin in the ground state³. Here, Lys 296 (helix VII:11 in the nomenclature used there³) is close to the centre of the membrane, and the β -ionone ring of retinal is close to the conserved tryptophan (at VI:16), with the side chain of tyrosine (at VI:19) which is found in all opsins, and the side chain of alanine (or threonine) (VI:20) found in mammalian opsins, below the retinal ring. This indicates that the β -ionone ring should point towards the intracellular side of the membrane, as shown here, but this intermolecular information for retinal with respect to the protein is not revealed by the current NMR data for symmetry reasons.

The unique structure of the chromophore reflects the remarkable efficiency of retinal, when located in the binding pocket of rhodopsin, to convert 30–35 kcal mol⁻¹ of light energy into a chemical signal by undergoing isomerization to an all-*trans* structure^{1,14,17}. The channelling of this energy to the protein body is thought to provide the main force for protein activation. To determine the orientation and proposed relaxed all-*trans* conformation of the chromophore in the M₁-state of the protein, ²H NMR experiments, were carried out on irradiated rhodopsin samples trapped at the M₁ state. For all C–C(²H)₃ bond vectors, angles between 60° and 65° were obtained (Table 1) reflecting the relaxed structure of the chromophore in agreement with the recently obtained torsion angle of 180° for the C₁₀–C₁₁ segment¹⁸, and intraligand distance measurements¹⁷ around this segment confirming a relaxed all-*E* structure. The structure shown in Fig. 3 indicates a combined movement of the ring and the C₆–C₁₁ segment relative to the Schiff base upon activation. The three bond vectors are roughly collinear, so the torsion angle of the chromophore plane relative to the membrane plane can only be accounted for by a tilt of the retinal axis. Using a reasonable value between 30° and 35° for the tilt of the axis (a change of ~10° occurs upon activation^{13,14}) a rotation of the chromophore plane through an angle of 45° to 50° occurs upon isomerization.

Deriving structurally relevant information at the atomic level for agonists and antagonists for GPCRs will enable detailed insights into their binding and activation, as shown here on the photo-receptor protein bovine rhodopsin. The structural information obtained allows a critical evaluation of binding sites and binding modes of GPCR ligands, particularly with respect to recent assumptions about the existence of a binding pocket within the transmembrane domain which may be highly conserved throughout the whole family of GPCRs^{3–6,8}. To define this structural feature of receptor activation is not only important for rhodopsin related diseases, but is of general relevance to the activation of the large family seven transmembrane domain GPCRs that rely on diffusing transmitter molecules. □

Methods

Sample preparation

Three different retinals were synthesized^{17,19}, all analogues of the naturally occurring chromophore but with each of the following methyl groups, C₅–C₁₈, C₉–C₁₉ and C₁₃–C₂₀, carrying deuterium C–C(²H)₃. Rhodopsin was isolated from bovine retina but with its indigenous retinal removed to give chromophore-free opsin^{17,19}. Each of the labelled retinals was introduced into the opsin to give a fully functional rhodopsin which is embedded in bilayers using well established methods^{19,20}.

NMR experiments

NMR experiments were carried out at 61.402 MHz on a BRUKER MSL 400 using a 7-mm MAS probe. Spectra were recorded using a 'single pulse' sequence with the sample subject

to MAS. The pulse length was typically 5 μ s. The spinning rate was stabilized to about 3 Hz. Measurements were performed at 213 K using a spinning speed of 2860 Hz, a recycle delay of 300 ms and around 400,000 scans. Spectra were processed with 16 K zero-filling applying 100 Hz of exponential line broadening.

Data analysis

The symmetrized NMR spectra were analysed using a program for MAS lineshape simulations of oriented systems based on the NMR software library GAMMA as described⁷. The observed MAS sideband distribution depends on the tilt angle β_{PM} of the C–C(²H)₃ bond vector with respect to the protein long axis Z_M, defined by the orientation of helix 7. The distribution of Z_M about the sample director Z_D, which is caused by some remaining disorder in the sample is described by a three-dimensional Gaussian distribution of the spectral width $\Delta\beta$. A spectral deconvolution analysis applied to spectrum of Fig. 2b showed that more than 80% of the sample was in the M₁ state.

To relate the orientational constraints obtained by the simulation package (obtained with respect to the membrane normal) to the protein body itself, a protein reference system was chosen based on the latest structure² for the arrangement and positions of the seven transmembrane helices in bovine rhodopsin in the ground state. The projection maps reveal helix 7, where 11-*cis* retinal is attached to the ϵ -amino group of Lys 296, as collinear with the membrane normal, and therefore helix 7 has been used here as representative of the protein long axis Z_M (Fig. 3).

Received 20 October 1999; accepted 17 March 2000.

- Birge, R. R. Nature of the primary photochemical events in rhodopsin and bacteriorhodopsin. *Biochim. Biophys. Acta* **1016**, 293–327 (1990).
- Krebs, A., Villa, C., Edwards, P. C. & Schertler, G. F. X. Characterisation of an improved two-dimensional p22₁₂₁ crystal from bovine rhodopsin. *J. Mol. Biol.* **282**, 991–1003 (1998).
- Baldwin, J. M., Schertler, G. F. X. & Unger, V. M. An alpha-carbon template for the transmembrane helices in the rhodopsin family of G-protein-coupled receptors. *J. Mol. Biol.* **272**, 144–164 (1997).
- Baldwin, J. M. Structure and function of receptors coupled to G proteins. *Curr. Opin. Cell Biol.* **6**, 180–190 (1994).
- Donnelly, D. & Findlay, J. B. C. Seven-helix receptors: structure and modelling. *Curr. Opin. Struct. Biol.* **4**, 582–589 (1994).
- Herzyk, P. & Hubbard, R. E. Combined biophysical and biochemical information confirms arrangement of transmembrane helices visible from the three-dimensional map of frog rhodopsin. *J. Mol. Biol.* **281**, 741–754 (1998).
- Glaubitz, C., Burnett, L. J., Gröbner, G., Mason, A. J. & Watts, A. Deuterium-MAS NMR spectroscopy on oriented membrane proteins: applications to photointermediates of bacteriorhodopsin. *J. Am. Chem. Soc.* **121**, 5787 (1999).
- Ji, T. H., Grossmann, M. & Ji, L. G. Protein-coupled receptors. *J. Biol. Chem.* **273**, 17299–17302 (1998).
- Altenbach, C., Cai, K., Khorana, H. G. & Hubbell, W. L. Structural features and light-dependent changes in the sequence 306–322 extending from helix VII to the palmitoylation sites in rhodopsin: a site-directed spin-labeling study. *Biochemistry* **38**, 7931–7937 (1999).
- Tan, Q. *et al.* Absolute sense of twist of the C12–C13 bond of the retinal chromophore in bovine rhodopsin based on exciton-coupled CD spectra of 11,12-dihydroretinal analogues. *Angew. Chem. Int. Ed.* **36**, 2089–2093 (1997).
- Buss, V., Kolster, K., Terstegen, F. & Vahrenhorst, R. Absolute sense of twist of the C12–C13 bond of the retinal chromophore in rhodopsin-semiempirical and nonempirical calculations of chiroptical data. *Angew. Chem. Int. Ed.* **37**, 1893–1895 (1998).
- Rothschild, K. J., Sanches, R., Hsiao, T. L. & Clark, N. A. A spectroscopic study of rhodopsin alpha-helix orientation. *Biophys. J.* **31**, 53–64 (1980).
- Jaeger, S. *et al.* Chromophore structural changes in rhodopsin form nanoseconds to microseconds following pigment photolysis. *Proc. Natl Acad. Sci. USA* **94**, 8557–8562 (1997).
- Shieh, T., Han, M., Sakmar, T. P. & Smith, S. O. The steric trigger in rhodopsin activation. *J. Mol. Biol.* **269**, 373–384 (1997).
- Eilers, M., Reeves, P. J., Ying, W., Khorana, H. G. & Smith, S. O. Magic angle spinning NMR of the protonated retinylidene Schiff base nitrogen in rhodopsin: Expression of ¹⁵N-lysine- and ¹³C-glycine-labeled opsin in a stable cell line. *Proc. Natl Acad. Sci. USA* **96**, 487–492 (1999).
- Feng, X. *et al.* Direct determination of a molecular torsional angle in the membrane protein rhodopsin by solid-state NMR. *J. Am. Chem. Soc.* **119**, 6853–6857 (1997).
- Verdegem, P. J. E. *et al.* Retinylidene ligand structure in bovine rhodopsin, metarhodopsin-I, and 10-methylrhodopsin from internuclear distance measurements using ¹³C-labeling and 1D rotational resonance MAS NMR. *Biochemistry* **33**, 11316–11324 (1999).
- Feng, X. *et al.* Determination of a molecular torsional angle in the metarhodopsin-I photointermediate of rhodopsin by double-quantum solid-state NMR. *J. Biomol. NMR* **16**, 1–8 (2000).
- Gröbner, G. *et al.* Photoreceptor rhodopsin: structural and conformational study of its chromophore 11-*cis* retinal in oriented membranes by deuterium solid state NMR. *FEBS Lett.* **422**, 201–204 (1998).
- Gröbner, G. *et al.* Macroscopic orientation of natural and model membranes for structural studies. *Anal. Biochem.* **254**, 132–138 (1997).
- Smith, S. O. *et al.* Low-temperature solid-state ¹³C NMR studies of the retinal chromophore in rhodopsin. *Biochemistry* **26**, 1606–1611 (1987).
- Hudson, B. S. & Birge, R. R. Angular orientation of the retinyl chromophore of bacteriorhodopsin: Reconciliation of ²H NMR and optical measurements. *J. Phys. Chem. A* **103**, 2274–2281 (1999).

Acknowledgements

We thank P. Fisher and B. Bonev for help, and J. Lugtenburg for discussion and advice. This work was supported by grants to A.W. (MRC, BBSRC, EU TMR, EU Biotech, Magnex, Varian, Bruker and HEFCE); C.G. is supported by a DFG Emmy Noether Fellowship.

Correspondence and requests for materials should be addressed to A.W. (e-mail: awatts@bioch.ox.ac.uk). The structural data for the chromophore can be obtained from the authors and will be posted at <http://www.bioch.ox.ac.uk/~awatts>.

Atomically defined mechanism for proton transfer to a buried redox centre in a protein

Kaisheng Chen^{*†}, Judy Hirst^{†‡§}, Raul Camba^{†‡}, Christopher A. Bonagura^{*}, C. David Stout^{||}, Barbara K. Burgess^{*} & Fraser A. Armstrong[‡]

^{*} Department of Molecular Biology and Biochemistry, University of California, Irvine, California 92612, USA

[‡] Department of Chemistry, Oxford University, Oxford OX1 3QR, UK

^{||} Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037-1083, USA

[†] These authors contributed equally to this work

[§] Present address: Medical Research Council Dunn Human Nutrition Unit, Hills Road, Cambridge CB2 2XY, UK

The basis of the chemiosmotic theory is that energy from light or respiration is used to generate a trans-membrane proton gradient¹. This is largely achieved by membrane-spanning enzymes known as 'proton pumps'^{2–5}. There is intense interest in experiments which reveal, at the molecular level, how protons are drawn through proteins^{6–13}. Here we report the mechanism, at atomic resolution, for a single long-range electron-coupled proton transfer. In *Azotobacter vinelandii* ferredoxin I, reduction of a buried iron–sulphur cluster draws in a solvent proton, whereas re-oxidation is 'gated' by proton release to the solvent. Studies of this 'proton-transferring module' by fast-scan protein film voltammetry, high-resolution crystallography, site-directed mutagenesis and molecular dynamics, reveal that proton transfer is exquisitely sensitive to the position and pK of a single amino acid. The proton is delivered through the protein matrix by rapid penetrative excursions of the side-chain carboxylate of a surface

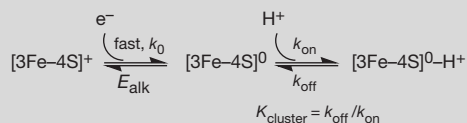
residue (Asp 15), whose pK shifts in response to the electrostatic charge on the iron–sulphur cluster. Our analysis defines the structural, dynamic and energetic requirements for proton courier groups in redox-driven proton-pumping enzymes.

Current studies of proton-pumping enzymes such as bacteriorhodopsin and cytochrome *c* oxidase are revealing the existence of 'proton wires' comprising chains of water molecules and protonatable amino-acid side-chains^{2–12}. Apart from water channels, which provide natural proton conductors¹⁴, transfer of protons is heavily restricted by their short tunnelling distance; thus, whereas electrons may easily tunnel 10 Å, a proton with comparable energy is limited to hops of less than 0.25 Å^{15–17}. This restriction can be used, along with barriers imposed by pK differences between proton donors and acceptors¹⁸, to control proton flow in response to electron-transfer events at nearby redox sites^{2–12}. However, proton pumps are complex membrane-bound enzymes and mechanistic details are difficult to analyse coherently. We now report conclusive studies of the mechanism of redox-driven proton transfer between solvent and a buried iron–sulphur cluster ([3Fe–4S]) in ferredoxin I (FdI) from *A. vinelandii*. This small protein provides an electron/proton-transfer 'module' with which to understand the redox-linked proton-transferring components of the larger enzymes^{13,19}. A major advantage of the FdI system is that native and mutant structures are known to high structural resolution in all

Box 1

Sequential electron and proton transfers at the [3Fe–4S] cluster in ferredoxin I

Scheme I. Sequence of electron and proton transfers defining the redox chemistry of the [3Fe–4S] cluster in Ferredoxin I. Electron transfers (standard first-order electrochemical rate constant, k_0) are fast, and E_{alk} is the reduction potential if $\text{pH} \gg \text{pK}_{\text{cluster}}$.



Scheme II. Sequence by which proton transfer to the cluster is catalysed by Asp 15 (B). Fast proton transfer (species highlighted in blue) is pH dependent, and protonation constants of Asp 15 are sensitive to cluster charge. At low pH, Asp 15 re-protonates (K_2), thus inhibiting proton transfer off the cluster. For native *A. vinelandii* FdI, $\text{pK}_{\text{OX}} = 5.4$. See Table I for rate expressions.

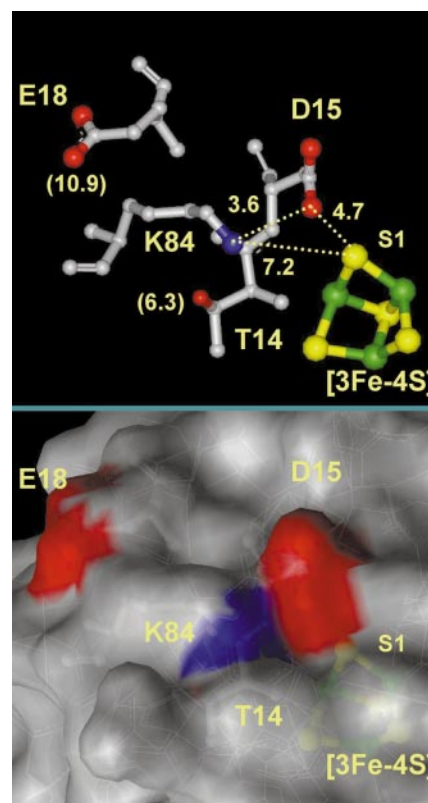
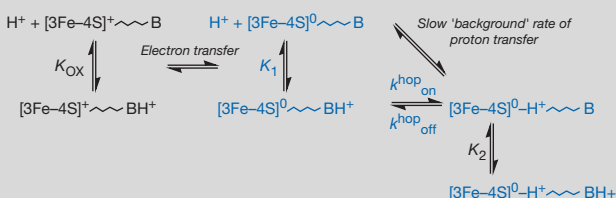


Figure 1 The residues in the vicinity of the [3Fe–4S] cluster in oxidized native *Azotobacter vinelandii* Ferredoxin I, taken from coordinates 7FD1. The upper frame shows the side chains of residues that were varied in this study, and indicates the separation (in Å) between D15 Oδ₁ and S1, the least buried μ₂ sulphide of [3Fe–4S]. Only the major conformation of the K84 side chain is shown (68% occupancy). The lower frame shows the same view, onto which is superimposed the solvent-accessible surface (calculated with a 1.4 Å probe). The carboxylate Oδ atoms of D15 and E18, and Nζ of K84 are solvent exposed. The Lys 84 side chain adopts two conformations in the oxidized native protein, as shown by the pH 8.5 structure (Protein data bank (PDB) code 7FD1, ref. 22) where these have 68% and 32% relative occupancies and (K84)Nζ–(D15)Oδ₁ distances of 3.6 Å and 3.1 Å respectively. The Glu 18 side chain is disordered (anomalously high B values).

relevant oxidation and protonation states^{20–22}. Importantly, FdI can be studied by fast-scan protein film voltammetry, a technique able to reveal, in detail, the bi-directional kinetics of coupled proton–electron-transfer reactions and their relation to thermodynamics^{13,23} (See Supplementary Information).

Scheme I (Box 1) shows the bi-directional proton transfer that accompanies fast electron transfer to the $[3\text{Fe}-4\text{S}]^{1+/0}$ cluster of FdI. Reduction drives a proton onto the cluster (k_{on}), while its reoxidation is ‘gated’ by proton release (k_{off}). Scheme I stems from numerous lines of evidence. The $[3\text{Fe}-4\text{S}]^{1+/0}$ reduction potential is pH-dependent, and one proton is taken up in the reduced state ($\text{pK}_{\text{cluster}} = 7.8$)^{13,23}. This uptake can be observed spectroscopically: thus, circular dichroism and magnetic circular dichroism spectra of oxidized FdI (spin-state (S) = 1/2) are independent of pH, whereas the one-electron reduced FdI ($[3\text{Fe}-4\text{S}]^0$, $S = 2$) shows an acid–base transition²⁴. The X-ray structures of oxidized and reduced FdI at high and low pH eliminate the possibility that the pH-dependent spectral changes are due to ligand exchange and/or structural rearrangements²⁰. Site-directed mutagenesis has established that the spectral changes are not due to protonation of the nearest ionizable residue (Asp 15)¹⁹, while Mössbauer spectroscopy has revealed that protonation perturbs the electronic structure of the

cluster²⁵. The obvious protonation site is a μ_2 sulphide (one of the three cluster sulphur atoms with a free coordination site).

Figure 1 shows the region above the buried $[3\text{Fe}-4\text{S}]$ cluster, both within the protein and looking down on the surface. Even at highest resolution (1.4 Å), the crystal structures, and NMR on a related protein²⁶, reveal no internal water molecules (or accommodating spaces) to act as proton-transfer agents^{14,20–22}. The starting point for this study is the mutant (D15N) in which Asp 15, the closest ionizable residue, which moves upon cluster reduction²⁰, is changed to Asn¹⁹. Protein film voltammetry showed that proton coupling is severely retarded in D15N FdI, indicating that the carboxylate acts as a proton relay group^{13,19}. However, in the native protein, Asp 15 is salt-bridged to Lys 84, raising the possibility of proton migration across the bridge. In addition, the highly mobile side chain of a second surface carboxylate residue, Glu 18, lies close by.

With the ability to make extensive kinetic, thermodynamic and structural measurements, we designed several new mutants. Figure 2 shows the arrangements of cluster-region amino acids for all variants, with distances between relevant atoms and the μ_2 sulphide (S1) situated closest to the protein surface. Circular dichroism spectroscopy confirmed that protonation of $[3\text{Fe}-4\text{S}]^0$ occurs in all cases. Table 1 displays the corresponding kinetic and

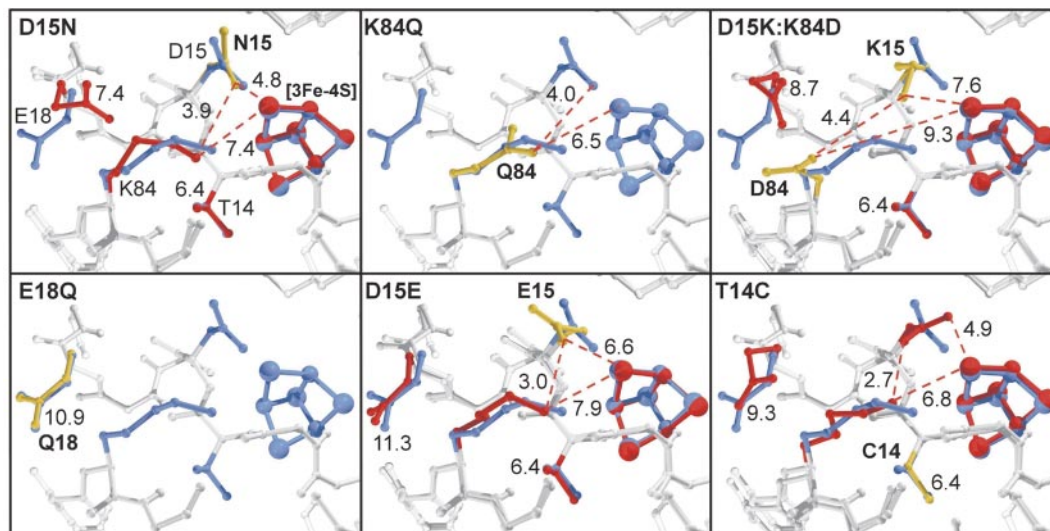


Figure 2 Structures of the various mutant forms of FdI in the region of interest, compared with the structure of oxidized native FdI. Side chains for the native protein are shown in blue and those altered by mutagenesis in yellow. Where X-ray structures are available, the positions of the other residues in the mutant proteins are shown in red. All distances are given in angstroms and where lines are not shown the number refers to the distance between the indicated residue and S1. Structures have been solved for D15N (PDB code

1FDD, ref. 19), D15K/K84D (PDB code 1B0T), D15E (PDB code 1D3W) and T14C (PDB code 1A6L, ref. 27). Crystals have not been obtained for K84Q or E18Q: these structures were modelled using the Insight II/Biopolymer program from Molecular Simulations for which the native oxidized structure (1.35 Å, PDB code 7FD1) was substituted with the desired mutant residue at the program-assigned minimum energy orientation.

Table 1 Kinetic and thermodynamic parameters for ‘on’ and ‘off’ proton transfers

Fast	E_{alk} (V)	$\text{pK}_{\text{cluster}}$ (high pH)	$\text{pK}_{\text{cluster}}$ (low pH)	$k_{\text{on}}(\text{M}^{-1}\text{s}^{-1})^*$ (pH 7.0)	$k_{\text{off}}(\text{s}^{-1})^*$ (pH 7.0)	pK_1	pK_2	$k_{\text{on}}^{\text{hop}}(\text{s}^{-1})$	$k_{\text{off}}^{\text{hop}}(\text{s}^{-1})$
Native	−0.443	7.8 ± 0.1	6.5 ± 0.1	7.9×10^9	308	7.2 ± 0.1	5.9 ± 0.1	$1,294 \pm 100$	332 ± 25
E18Q	−0.453	7.7 ± 0.1	6.7 ± 0.1	4.8×10^9	207	7.1 ± 0.1	6.1 ± 0.1	930 ± 90	230 ± 25
T14C	−0.464	8.4 ± 0.1	7.1 ± 0.1	6.6×10^9	207	8.0 ± 0.1	6.7 ± 0.1	720 ± 70	310 ± 25
K84Q	−0.476	8.1 ± 0.1	6.6 ± 0.1	9.0×10^9	232	7.4 ± 0.1	5.9 ± 0.1	$1,252 \pm 100$	250 ± 25
Native (D_2O)	−0.443	7.8 ± 0.1	6.5 ± 0.1	6.0×10^9	222	7.2 ± 0.1	5.9 ± 0.1	970 ± 100	240 ± 25
Slow	E_{alk} (V)	$\text{pK}_{\text{cluster}}$		$k_{\text{on}}(\text{M}^{-1}\text{s}^{-1})$	$k_{\text{off}}(\text{s}^{-1})$				
D15N	−0.408	6.9 ± 0.1		2.0×10^7	2.5 ± 0.1				
D15K–K84D	−0.397	6.6 ± 0.1		1.2×10^7	3.0 ± 0.1				
D15E	−0.388	6.7 ± 0.1		2.0×10^7	4.5 ± 0.2				

All terms are as defined in Schemes I and II.

*For the fast reactions, interpreted in terms of Scheme II, $k_{\text{on}} = k_{\text{on}}^{\text{hop}}/([\text{H}^+] + K_1)$ and $k_{\text{off}} = k_{\text{off}}^{\text{hop}}/([\text{H}^+] + K_2)$. Agreement with the simple bimolecular rate law of Scheme I requires consideration of the fact that interaction with Asp 15 causes the pK of the cluster to differ at high and low pH. In all cases k_{on} , the standard first-order electrochemical rate constant for electron exchange at the reduction potential, is $> 200 \text{ s}^{-1}$; the exponential increase in rate that occurs as a driving force is applied means that electron transfer is never rate limiting.

thermodynamic parameters for 'on' and 'off' proton transfers, defined according to Schemes I and II (Box 1), and measured by protein film cyclic voltammetry. In all cases, reduction involves electron transfer followed by proton transfer, whereas oxidation requires that proton transfer precede electron transfer, that is, the two processes do not occur in concert. Electron transfer is very fast and is clearly separated from proton transfer.

The mutants fall sharply into two categories with respect to proton-transfer kinetics—'slow' and 'fast'—with rates differing by three orders of magnitude. The kinetics of the 'slow' mutants are simple and pH-independent (they are described adequately by Scheme I) and rates are similar in each case. By contrast, the 'fast' mutants show a complex pH dependence (Scheme II) with second-order 'on' rate constants approaching diffusion control at pH 7.

Referring to Fig. 2, the following facts are established. The D15N mutation replaces the carboxylate by carbamide and the slow proton-transfer rate is consistent with loss of a protonatable group (base B) which acts as a proton relay (Scheme II)^{13,19}. However, the mutation also disrupts the salt bridge between the Asp 15 carboxylate and a surface lysine (Lys 84)¹⁹. In addition, the side chain of Glu 18 shifts, although the change is less-defined due to disorder. In the mutant K84Q, Lys 84 is changed to a residue (Gln) that cannot form a salt-bridge to Asp 15. The fast proton-transfer kinetics demonstrate that it is the presence of the carboxylate and not the salt-bridge that is important. The double mutant D15K/K84D was designed to invert the salt-bridge orientation; however, the salt-bridge does not form. Proton transfer is slow, as it is in D15N, showing that the introduced Asp 84 side chain is incorrectly placed to function as a proton-transfer group and that lysine cannot substitute for aspartate at position 15. The possibility that Glu 18 facilitates proton transfer was eliminated by the mutant E18Q, which has very similar rates to the native protein. Experiments with D15E finally established the critical nature of the distance between the cluster and the carboxylate at position 15. Insertion of a single CH₂ group in the side chain (which increases the distance by approximately 2 Å) retards proton transfer as much as deletion of the carboxylate altogether (Table I). The T14C mutant has a polarizable S atom within the sphere of influence of the cluster and raises the cluster pK while the position of the carboxylate is not significantly changed²⁷. The native proton-transfer kinetics are retained, confirming that they are dependent solely on the position of the carboxylate.

The exacting requirement for Asp 15 is thus demonstrated, leading to a detailed model for the mechanism of proton transfer between water at the protein surface and the buried redox centre. The data in Table 1 show that the H₂O/D₂O isotope effect for the native protein is small (approximately 1.3), and reveal (from the interdependence of respective pK values) that there are significant electrostatic interactions between the cluster and Asp 15. As shown in Fig. 3a, electron transfer drives rotation of the Asp 15 carboxylate about the Cβ–Cγ bond and increases the Oδ to S1 distance from 4.7 Å to 4.9 Å, in accordance with the less favourable electrostatics²². More significantly, the pK of the carboxylate increases to 7.2 compared with 5.4 in the oxidized ([3Fe–4S]¹⁺) state (pK_{OX} in Scheme II), thus promoting proton capture from solvent water¹⁸. It is important to note that a sizeable, transient shift in pK value is essential for fitting the data. After the proton has transferred to the cluster, the pK drops back to 5.9, and the X-ray structure of the reduced protein at pH 6.1 reveals that the Asp 15 carboxylate reverts towards its position in the oxidized form²⁰. This 'relaxation' is expected because the [3Fe–4S]⁰–H⁺ cluster has the same electrical charge as the oxidized [3Fe–4S]¹⁺ cluster.

To explain how Asp 15 mediates such fast proton transfer, it is necessary to ascertain whether a proton that has transferred from solvent to a carboxylate O atom can then be carried easily to within hydrogen-bonding distance of the cluster S1 atom²⁸. We therefore carried out molecular dynamics calculations to study the mobility

of the state in which the reduced cluster is unprotonated while Asp 15 is protonated and thus no longer repelled by the increased negative charge on the cluster. The results reveal that the Asp 15 side chain is very mobile, executing a high frequency of short-range encounters between Oδ and S1 atoms (Fig. 3b). Thus, within 80 ps, an Oδ atom makes one excursion to 3.05 Å (Fig. 3c), well within the sum of van der Waals radii, and five other excursions to distances less than 4 Å. The motions of the side-chain carboxyl/carboxylate group thus convey H⁺ between solvent and reduced cluster, accomplishing 'atom-to-atom' transfer across this hydrophobic barrier. Cluster deprotonation occurs by reversal of these steps, noting that at pH < 5.9, protonation of Asp 15 inhibits proton transfer off the cluster.

In conclusion, fast proton transfer in FdI requires the presence of Asp 15 and rapid penetrative excursions of its side-chain carboxyl/carboxylate to within hydrogen-bonding distance of the cluster. Our experiments, which combine detailed kinetic and thermodynamic data for a series of mutants with molecular dynamics based on high-resolution protein crystal structures, highlight the exacting specifications that must be met for proton-pumping motifs in enzymes^{2–5}. The pK values of the proton-relaying carboxylate and the buried active site are tightly coupled and adjust to facilitate sequential transfers of an electron and a proton. Re-oxidation is gated by

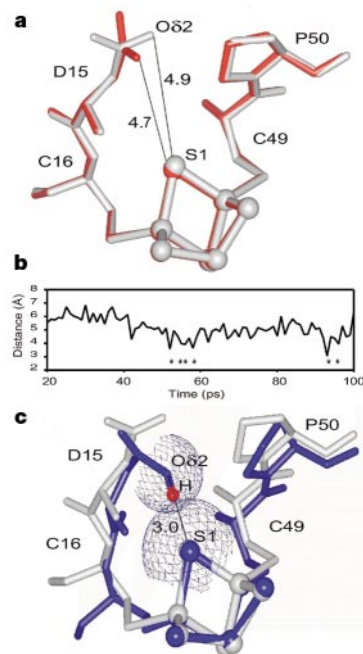


Figure 3 Movement of the Asp15 side chain during redox-driven proton transfers. **a**, Comparison of oxidized (red; PDB code 7FD1) and reduced (grey; PDB code 7FDR) high-pH structures of native FdI in the region of the cluster and Asp 15. Upon reduction, the Lys 84 side chain adopts a single conformation with a Nζ–Oδ₁ distance of 3.3 Å, and the carboxylate side chain has rotated by 90°. (In accordance with previous nomenclature, the closest O atom to the cluster in the reduced form is denoted Oδ₂, whereas it is Oδ₁ in the oxidized form²².) **b**, Distances of Oδ₂ (from Asp 15) to S1 (from the cluster) during the molecular dynamics simulation; asterisks indicate the points where the distance is below 4 Å. **c**, Superposition of the reduced native FdI (grey) structure (also shown in **a**) and one of the molecular-dynamics-generated snapshot structures (blue) around the region of Asp 15, showing one of the approaches of Oδ₂ to the S1 atom of [3Fe–4S]⁰ within the period of simulation. The distance between Oδ₂ and S1 is 4.9 Å for the native structure and 3.0 Å for the molecular-dynamics-generated structure. The H atom is positioned collinear with Oδ₂–S1 at a normal bond distance (1.0 Å) from Oδ₂. Van der Waals surfaces of Oδ₂ (radius 1.4 Å) and S1 (1.85 Å) are shown. For this closest approach, Oδ and S1 are well within the sum (1.4 + 1.85 = 3.25 Å) of their van der Waals radii and, in the limit of a collinear O–H...S arrangement, the resulting H–S distance is approximately 2.25 Å.

proton transfer, demonstrating how biology can control long-range electron transfer by linking it to a far more discriminating process. Mechanistic requirements for proton transfer are so stringent that even minor structural changes produce decreases in rate over orders of magnitude. □

Methods

Protein film voltammetry

Protein film voltammetry probes electron-transfer processes that are induced by perturbing the electrochemical potential applied to a mono-/submono-layer of protein molecules bound at an electrode surface (See Supplementary Information). Electron transfers in and out of the active site are observed as electrical current signals, which are altered in specific ways if there is coupling to processes such as proton transfer. Thus, by analysing signals over a range of voltage scan rate and pH, a detailed and integrated picture of the coupling kinetics and energetics is obtained. All measurements were carried out as described previously¹³ and 'trumpet plots' of peak positions versus scan rate were analysed in terms of the kinetics and thermodynamics described by Scheme II.

Site directed mutagenesis and protein crystallography

Site directed mutagenesis and protein purification were carried out as previously reported^{19,27}. Crystal structures were determined accordingly to procedures recently described²².

Molecular dynamics

Molecular dynamics simulations were carried out using the DISCOVER program from Molecular Simulations. The AMBER force field used for the protein and Fe-S cluster was as described²⁹ except that cluster atomic charges were derived from X α density functional calculations³⁰. The structure of reduced FdI at pH 8.5 (1.35 Å resolution, protein data bank code 7FDR, ref. 21) was used as the starting model except that the O δ_2 atom of Asp 15 was protonated. Besides the crystallographic water molecules, a 9 Å layer of water molecules was added around the protein and those occupying the outermost 5 Å were constrained to prevent their evaporation. After initial minimization, the system was heated for 5 ps at 100, 200, 273 K, and then the dynamics were run at 273 K for 100 ps with a time step of 1.5 fs: the first 20 ps were discarded. The mean structure over the 80 ps considered had r.m.s. deviation values from the starting structure of 1.20 Å for the backbone atoms and 1.45 Å for all heavy atoms. The individual structures generated had average r.m.s. deviation values (from the mean structure) of 0.68 Å for the backbone atoms and 0.86 Å for all heavy atoms, within the range expected²⁶.

Received 26 November 1999; accepted 12 April 2000.

- Nicholls, D. G. & Ferguson, S. J. Bioenergetics 2 (Academic, San Diego, 1992).
- Wikström, M. Proton translocation by bacteriorhodopsin and heme-copper oxidases. *Curr. Opin. Struct. Biol.* **8**, 480–488 (1998).
- Gennis, R. B. How does cytochrome oxidase pump protons? *Proc. Natl Acad. Sci. USA* **95**, 12747–12749 (1998).
- Malmström, B. G. Cytochrome oxidase: pathways for electron tunneling and proton transfer. *J. Biol. Inorg. Chem.* **3**, 339–343 (1998).
- Michel, H. The mechanism of proton pumping by cytochrome *c* oxidase. *Proc. Natl Acad. Sci. USA* **95**, 12819–12824 (1998).
- Rammelsberg, R., Huhn, G., Lübben, M. & Gerwert, K. Bacteriorhodopsin's intramolecular proton-release pathway consists of a hydrogen-bonded network. *Biochemistry* **37**, 5001–5009 (1998).
- Luecke, H., Schobert, B., Richter, H.-T., Cartailler, J. P. & Lanyi, J. Structural changes in bacteriorhodopsin during ion transport at 2 Å resolution. *Science* **286**, 255–260 (1999).
- Nabedryk, E., Breton, J., Okamura, M. Y. & Paddock, M. L. Proton uptake by carboxylic acid groups upon photoreduction of the secondary quinone (Q_B) in bacterial reaction centres from *Rhodospirillum rubrum*: FTIR studies on the effects of replacing Glu H173. *Biochemistry* **37**, 14457–14462 (1998).
- Yoshikawa, S. *et al.* Redox-coupled crystal structural changes in bovine heart cytochrome *c* oxidase. *Science* **280**, 1723–1729 (1998).

- Konstantinov, A. A., Siletsky, S., Mitchell, D., Kaulen, A. & Gennis, R. B. The roles of two proton input channels in cytochrome *c* oxidase from *Rhodospirillum rubrum* probed by the effects of site-directed mutations on time-resolved electrogenic intraprotein proton transfer. *Proc. Natl Acad. Sci. USA* **94**, 9085–9090 (1997).
- Lübben, M., Prutsch, A., Mamat, B. & Gerwert, K. Electron transfer induces side-chain conformational changes of glutamate-286 from cytochrome *bo*₃. *Biochemistry* **38**, 2048–2056 (1999).
- Junemann, S., Meunier, B., Fisher, N. & Rich, P. R. Effects of mutation of the conserved glutamic acid-286 in subunit I of cytochrome *c* oxidase from *Rhodospirillum rubrum*. *Biochemistry* **38**, 5248–5255 (1999).
- Hirst, J. *et al.* Kinetics and mechanism of redox-coupled, long-range proton transfer in an iron-sulfur protein. Investigation by fast-scan protein-film voltammetry. *J. Am. Chem. Soc.* **120**, 7085–7094 (1998).
- Meyer, E. Internal water molecules and H-bonding in biological macromolecules: a review of structural features with functional implications. *Protein Sci.* **1**, 1543–1562 (1992).
- Beratan, D. N., Onuchic, J. N., Winkler, J. R. & Gray, H. B. Electron-tunneling pathways in proteins. *Science* **258**, 1740–1741 (1992).
- Page, C. G., Moser, C. C., Chen, X. & Dutton, P. L. Natural engineering principles of electron tunnelling in biological oxidation-reduction. *Nature* **402**, 47–52 (1999).
- Klinman, J. P. Quantum mechanical effects in enzyme-catalyzed hydrogen transfer reactions. *Trends Biochem. Sci.* **14**, 368–373 (1989).
- Gutman, M. & Nachliel, E. The dynamics of proton exchange between bulk and surface groups. *Biochim. Biophys. Acta* **1231**, 123–138 (1995).
- Shen, B. *et al.* Azotobacter vinelandii ferredoxin I. Aspartate 15 facilitates proton transfer to the reduced [3Fe-4S] cluster. *J. Biol. Chem.* **268**, 25928–25939 (1993).
- Stout, C. D. Crystal structures of oxidized and reduced Azotobacter vinelandii ferredoxin at pH 8 and 6. *J. Biol. Chem.* **268**, 25920–25927 (1993).
- Stout, C. D., Stura, E. A. & McRee, D. E. Structure of Azotobacter vinelandii 7Fe ferredoxin at 1.35 Å resolution and determination of the [Fe-S] bonds with 0.01 Å accuracy. *J. Mol. Biol.* **278**, 629–639 (1998).
- Schipke, C. G., Goodin, D. B., McRee, D. E. & Stout, C. D. Oxidized and reduced Azotobacter vinelandii ferredoxin I at 1.4 Å resolution: conformational change of surface residues without significant change in the [3Fe-4S]^{+/0} cluster. *Biochemistry* **38**, 8228–8239 (1999).
- Armstrong, F. A., Heering, H. A. & Hirst, J. Reactions of complex metalloproteins studied by protein-film voltammetry. *Chem. Soc. Rev.* **26**, 169–179 (1997).
- Stephens, P. J. *et al.* Circular-dichroism and magnetic circular-dichroism of Azotobacter vinelandii ferredoxin I. *Biochemistry* **30**, 3200–3209 (1991).
- Hu, Z. G., Jollie, D., Burgess, B. K., Stephens, P. J. & Münch, E. Mössbauer and EPR studies of Azotobacter vinelandii ferredoxin I. *Biochemistry* **33**, 14475–14485 (1994).
- Aono, S. *et al.* Solution structure of oxidized Fe₂S₂ ferredoxin from the thermophilic bacterium Bacillus schlegelii by ¹H NMR spectroscopy. *Biochemistry* **37**, 9812–9826 (1998).
- Gao-Sheridan, A. T14C variant of Azotobacter vinelandii ferredoxin I undergoes facile [3Fe-4S]⁰ to [4Fe-4S]²⁺ conversion in vitro but not in vivo. *J. Chem. Biol.* **273**, 33692–33701 (1998).
- Guthrie, J. P. Intrinsic barriers for protons transfer reactions involving electronegative atoms, and the water mediated proton switch: an analysis in terms of Marcus theory. *J. Am. Chem. Soc.* **118**, 12886–12890 (1996).
- Banci, L., Bertini, I., Carloni, P., Luchinat, C. & Orioli, P. L. Molecular-dynamics on HIPIP from Chromatium vinosum and comparison with NMR data. *J. Am. Chem. Soc.* **114**, 10683–10689 (1992).
- Noodleman, L., Norman, J. G. Jr, Osborne, J. H., Aizman, A. & Case, D. A. Models for ferredoxins—electronic structures of iron sulfur clusters with one, two and four iron atoms. *J. Am. Chem. Soc.* **107**, 3418–3426 (1985).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

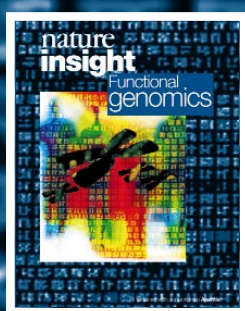
Acknowledgements

We thank T. Poulos and J. Lanyi for comments on the manuscript. This research was supported by grants from the NIH, EPSRC, and BBRSC. B.B.K. thanks The Fulbright Commission for a Senior Scholarship, and the John Simon Guggenheim Foundation for a Travelling Fellowship. R.C. is grateful to The National Council of Science and Technology of Mexico (CONACYT) for their support.

Correspondence and requests for materials should be addressed to F.A.A. (e-mail: fraser.armstrong@chem.ox.ac.uk).

nature insight

Functional genomics



Cover illustration
courtesy of Jacey.

Functional genomics has leapt from being a surrealistic, or at least futuristic, concept in the 1980s to an accepted (if not yet everyday) part of science in the year 2000. How has this transformation come about? Since worldwide efforts to sequence genomes began formally in 1990, astounding technological advances have been introduced. Among the eukaryotes, yeast, worm and fly sequences have been completed, alongside more than 20 prokaryotic genomes. The expected date for completion of the entire human genome is 2003, with a first draft due this autumn.

But what is the value of all this sequence data? An inventory of genes will impact molecular medicine the greatest, leading to improved diagnosis of disease. Sequencing of prokaryotic genomes will aid vaccine design and allow exploration of new microbial energy sources, while knowledge of other animal and plant genomes should enhance agriculture. Gaining the DNA sequences heralds the end of the beginning. The next step in this biological revolution is 'functional genomics', not simply the assignation of function to the identified genes but the organization and control of genetic pathways that come together to make up the physiology of an organism. This month's *Nature Insight* focuses on the challenges to biology brought about by the avalanche of DNA sequence information.

Vukmirovic and Tilghman provide an overview to the genomic revolution on page 820 and discuss what it will mean to scientists interested in the fundamentals of life. The progression of biology into a data-rich science has been orchestrated by computational biologists. On page 823, David Eisenberg and colleagues look at the role computers will play in predicting the function of a gene and even modelling signalling pathways in which it may act. At the molecular level, functional information can be acquired through the analysis of DNA and RNA expression arrays and on page 827 Lockhart and Winzler examine the current status of this technology. On page 837, Pandey and Mann discuss the sophisticated machinery being used in proteomics — the large-scale analysis of proteins and their interactions. The past couple of decades have witnessed an explosion in the identification of genes for several inherited human disorders. But successes have been limited mainly to diseases caused by mutations in a single gene. Neil Risch discusses on page 847 how having the human genome at our fingertips will present new opportunities for geneticists studying complex human disorders. Finally, on page 857 Allen Roses introduces pharmacogenetics — the study of how genetic differences influence the variability in patient response to drugs and allow custom-drug design.

We are pleased to acknowledge the financial support of Aventis in producing this Insight. Of course, *Nature* carries the sole responsibility for all editorial content and rigorous peer-review. In 1953, *Nature* published the structure of the DNA helix. Today, as the first human chromosome sequences appear in our pages, we stand at the brink of the next biological revolution. We hope that our readers will find the following reviews enlightening as well as thought provoking. The sequence for the human chromosomes and published genomes can be accessed online through *Nature's* Genome Gateway at <http://www.nature.com/genomics>.

Ritu Dhand Insight Editor

Publisher and liaison for corporate support [Liz Allen \(e.allen@nature.com\)](mailto:Liz.Allen@nature.com)

Exploring genome space

Ognjenka Goga Vukmirovic & Shirley M. Tilghman

Howard Hughes Medical Institute and Department of Molecular Biology, Princeton University, Princeton New Jersey 08544, USA

The completion of entire genome sequences of many experimental organisms, and the promise that the human genome will be completed in the next year, find biology suddenly awash in genome-based data. Scientists are scrambling to develop new technologies that exploit genome data to ask entirely new kinds of questions about the complex nature of living cells.

Biology is in the midst of an intellectual and experimental sea change. Essentially the discipline is moving from being largely a data-poor science to becoming a data-rich science. The data are coming from the fortuitous confluence of technological advances in protein and DNA analysis as well as imaging advances in cell biology. Improvements in mass spectrometry have revolutionized the number and kind of proteins that can be identified in a cell, and the powerful tools of nuclear magnetic resonance spectroscopy and X-ray crystallography have been managing to keep pace with far more efficient methods for acquiring angstrom-level structural knowledge for individual proteins, as well as multi-protein complexes. Reports of the structure of the nucleosome¹ and the RNA polymerase complex² and the promise of a complete picture of the ribosome in the near future^{3,4} are landmark events in biology, and whet our appetite for more. At the same time, entire genome sequences of a large number of prokaryotes and a rapidly growing number of eukaryotes are now in hand. The exciting prospect of having a 90% draft of the human genome is almost at hand. The review articles that follow in this *Nature* Insight highlight some of the challenges that biologists face as they acclimatize themselves to this change in the data landscape.

Data drives innovation

That data are inherently good was not a central philosophical tenet for all biologists. This was evident in the late 1980s when the wisdom of embarking on the Human Genome Project was being debated in the community. At that time, many argued that investing in genome sequencing was unwise until we had the tools in hand to understand the sequence. Funds spent acquiring the sequence would be better spent developing tools to first understand it. To physicists and engineers, in disciplines where data are very often acquired well before their utility is apparent, this seemed illogical. The past ten years has provided ample evidence for the value of large-scale data acquisitiveness in biology. In fact, by thinking boldly, and by setting ambitious goals for itself, the international Human Genome Project stimulated developments both in high-throughput DNA sequencing, which were essential for the success of the project, and in powerful computational tools for sequence analysis. Without the project, it is very unlikely that these improvements would have been developed. The data were the catalyst.

Assembling the parts list

This avalanche of data is changing the kinds of questions that biologists can ask. Until recently, scientists have studied the form and function of organisms primarily by

narrowing their focus from the entire animal to increasingly smaller parts — first organs, then cells, and finally individual molecules. Essentially the enormous complexity of a living organism overwhelmed existing analytical tools, and real progress came from approaches that ignored the complexity and focused on the component parts. This was a powerful strategy, and one that will continue to be important. A metaphor that captures the reductionist approach is that of a child trying to understand the function of a mysterious black box by gradually taking it apart, and examining each part individually. Today the parts, the individual genes and proteins, are now on the table for a growing number of organisms. As in the case of the black box, some parts are

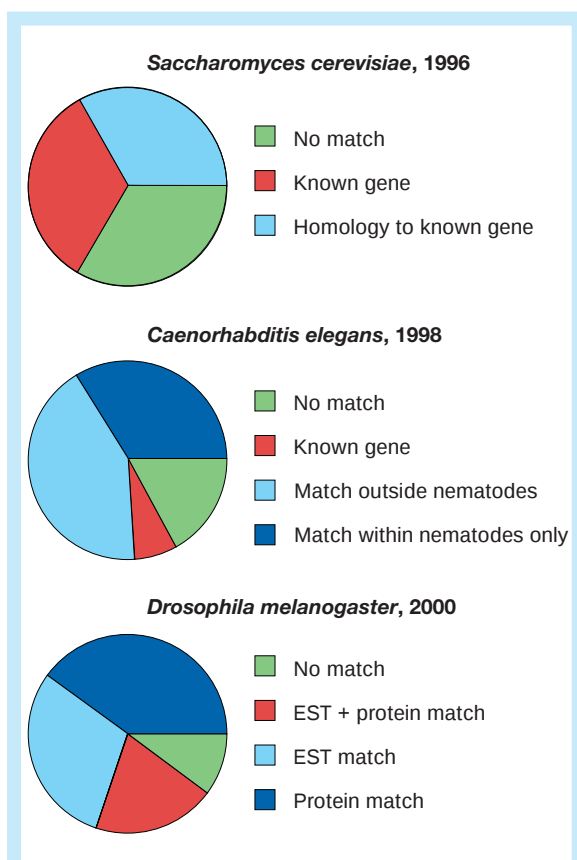


Figure 1 The distribution of genes in eukaryotic genomes. Shown for three organisms are the relative number of genes that were previously identified, the number that had some homology to known genes, and the number that had no match in any sequence database at the time of completion of the genome sequence.

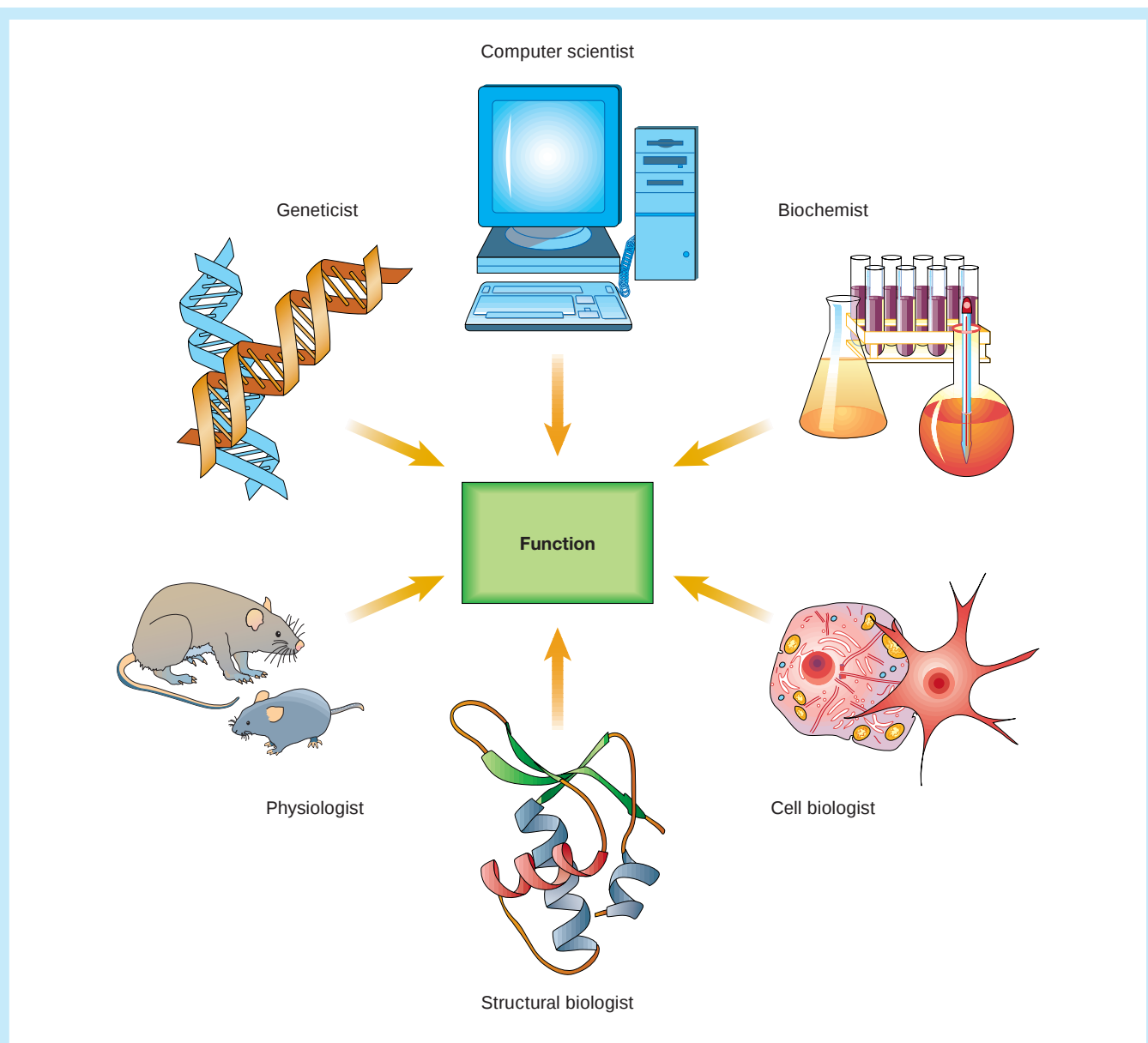


Figure 2 Understanding gene function. The function of a specific gene can be approached from many scientific perspectives with a variety of tools.

immediately recognizable because they have been encountered before, some can elicit an educated guess as to their function because of their similarity to another gene, and the rest are complete mysteries. When the *Saccharomyces cerevisiae* genome was published, it was estimated that of its ~6,000 genes, 2,000 had been studied previously, 2,000 bore some resemblance to known genes and 2,000 were unrecognizable (Fig. 1)⁵. The situation is even more striking in multicellular organisms. When the *Caenorhabditis elegans* genome was published in 1998, it was reported that of the ~19,000 genes, only 7% had been studied previously, although 42% of the genes had some match to proteins and sequences of random complementary DNAs (expressed sequence tags) from many organisms other than nematodes⁶. These matches can often be clues to the function of previously unstudied genes. By 2000, the number of completely novel genes with no match to anything previously encountered in DNA sequence was reduced to 17% of the 13,600 *Drosophila* genes in the fly genome⁷.

From sequence to function

Once whole-genome information is available for an organism, the challenge turns from identifying the parts to understanding their

function, thus ushering in the 'post-genomic' era, also referred to as 'functional genomics'. In the short term, the goal is to assign some element of function to each of the genes in an organism, and to do this with high-throughput, systematic approaches. With two-thirds of the yeast genome still to be assigned function, the notion of accumulating this information one gene at a time is hard to contemplate. This knowledge gap has been the crucial impetus for developing 'whole-genome' approaches that can acquire functional information, in the form of expression profiles (see review by Lockhart and Winzler, pages 827–836), protein–protein interactions (see review by Pandey and Mann, pages 837–846), computational approaches (see review by Eisenberg *et al.*, pages 823–826) and the response to loss of function by mutation, sometimes called genetic fingerprinting^{8,9}.

Function is understood on many different levels in biology. A computational biologist might be content to identify an unknown gene as encoding a kinase, but the biochemist would want to know its substrate specificity, the cell biologist would want to know its intracellular localization and its targets, the geneticist would want to know the pathway it affects and the physiologist would want to know

what organs it affects (Fig. 2). All of this information contributes to the sum of our understanding of function, and there is a critical demand for algorithms (for example, see ref. 10) and relational databases that can integrate the information that will be obtained using very different tools.

Several of the articles in this Insight discuss new technologies that are being developed to understand function on a genome-wide scale. Are these approaches conceptually different from what biologists have been doing for many years, or is it just the scale on which experiments can be done that is different? One could argue, for example, that geneticists who have been conducting screens to identify all genes in a pathway for 80 years are conducting a 'whole-genome' experiment. The major difference between the pre- and post-genome era is that one can now potentially account for and keep track of all the components at once. Theoretically there are no unknowns, with respect to genes and proteins. For example, Eisenberg *et al.*, in their review on pages 823–826, describe a comparative genomic study by Pellegrini *et al.*¹¹ to identify genes that are conserved *en bloc* in a variety of prokaryotes. From this analysis, they deduced the functions of unknown genes by the fact that they co-evolved with genes of known function — essentially guilt by association. It was the comprehensive nature of whole-genome sequences that made this a feasible approach to identifying function.

Thus the unique aspect of functional genomics in an organism whose genome is known completely is the ability to monitor simultaneously potentially all events, whether it be the expression of genes at the RNA or protein level, all possible protein–protein interactions, all alleles of all genes that affect a particular trait, or all protein-binding sites in a genome. The potential to learn entirely new things by taking such an approach is enormous, but it is important to recognize that these are early days for genomics. The recent developments in simultaneous monitoring of the expression of all RNAs using oligonucleotide and cDNA arrays, reviewed comprehensively by Lockhart and Winzler (pages 827–836), is a case in point. Although few would question the power of this technology to describe the transcriptional profile of a cell, critics point out that the computational tools to maximally extract new insights into biology are not yet in hand. This is no different from the situation at the beginning of genome sequencing described above, and its resolution will undoubtedly be the same. The very existence of these large data sets will stimulate the development of better analytical tools. The array technologies suggest hypotheses about gene function that will stimulate new experiments that will use a reductionist approach. Thus there will be constant cross-fertilization and interplay between genome-wide and focused studies.

The technological and conceptual bottlenecks are not restricted to gene expression arrays. If parallel developments in proteomics are to make their mark, they must be able to monitor the protein modifications that are critical for cellular regulation on a proteome-wide scale, an issue that is raised in the review by Pandey and Mann (pages

837–846). Risch has laid out in his review (pages 847–856) the considerable challenge that faces human geneticists who are trying to take whole-genome approaches to identifying genes that underlie multi-genic human traits. Although the completion of the human genome sequence holds the promise of identifying all genes associated with disease, the successes to date have been restricted primarily to genes that are responsible for diseases caused by mutations in a single gene. Solving this problem will have a considerable impact on the success of pharmacogenetics, a new field reviewed by Roses (pages 857–865), whose goal is to tailor drugs to individuals' genomic makeup.

From function to integration

These are indeed exciting times, not unlike the early days of recombinant DNA in the 1970s, in which a revolutionary new technology permitted entirely new questions about the nature of genes to be raised. Although the current research is focused on assigning function to genes and proteins, the long-term goal is just as it is for the child and the black box — that is, to be able to understand sufficiently well how the pieces work together that you could, in principle, put them back together and get a functional organism. The challenge is to describe the collective properties of whole organisms in a precise and quantitative way. This challenge is new to biology, and its resolution will require, in addition to existing paradigms of molecular biology, new sets of analytical tools. It is hardly a coincidence that many universities and research institutes, including our own, are making major investments in multidisciplinary life-science initiatives to explore the complexity of living things. Organisms are networks of genes, which make networks of proteins, which regulate genes, and so on *ad infinitum*. The amount of complex data that will be generated, and the need for modelling to understand the way networks function, will ensure that disciplines outside of biology will be required to collaborate on this problem, if the ultimate goal to deconstruct such networks is to come to fruition. □

1. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
2. Zhang, G. *et al.* Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. *Cell* **98**, 811–824 (1999).
3. Ban, N. *et al.* Placement of protein and RNA structures into a 5 Å-resolution map of the 50S ribosomal subunit. *Nature* **400**, 841–847 (1999).
4. Cate, J. H., Yusupov, M. M., Yusupova, G. Z., Earnest, T. N. & Noller, H. F. X-ray crystal structures of 70S ribosome functional complexes. *Science* **285**, 2095–2104 (1999).
5. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563–547 (1996).
6. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
7. Adams, M. D. *et al.* The genome sequence of *Drosophila melanogaster*. *Science* **287**, 2185–2195.
8. Liu, L. X., *et al.* High-throughput isolation of *Caenorhabditis elegans* deletion mutants. *Genome Res.* **9**, 859–867 (1999).
9. Smith, V., Chou, K. N., Lashkari, D., Botstein, D. & Brown, P. O. Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274**, 2069–2074 (1996).
10. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).
11. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96**, 4285–4288 (1999).

Protein function in the post-genomic era

David Eisenberg, Edward M. Marcotte, Ioannis Xenarios & Todd O. Yeates

Molecular Biology Institute and UCLA-DOE Laboratory of Structural Biology and Molecular Medicine, Box 951570, University of California at Los Angeles, Los Angeles, California 90095-1570, USA (e-mail: david@mbi.ucla.edu)

Faced with the avalanche of genomic sequences and data on messenger RNA expression, biological scientists are confronting a frightening prospect: piles of information but only flakes of knowledge. How can the thousands of sequences being determined and deposited, and the thousands of expression profiles being generated by the new array methods, be synthesized into useful knowledge? What form will this knowledge take? These are questions being addressed by scientists in the field known as 'functional genomics'.

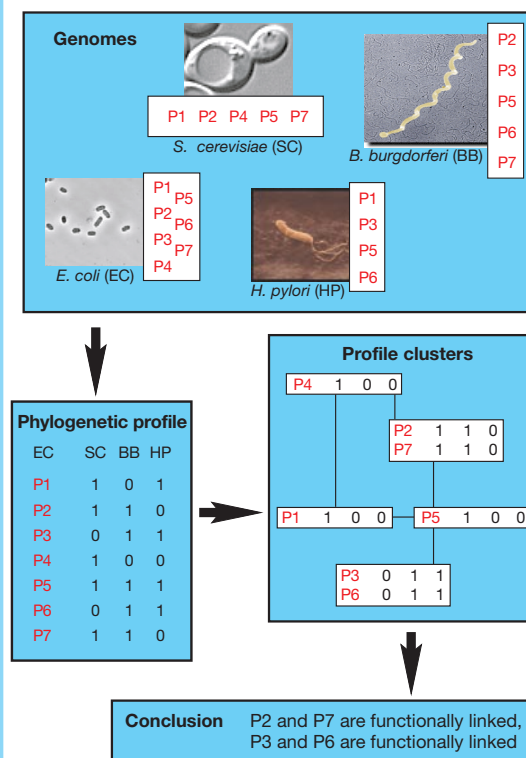
Inherent in the growing collections of genome sequences and expression profiles is knowledge about functional linkages between proteins. This knowledge can be extracted both by experimental and by computational means, as outlined below. New computational methods go beyond the traditional method of sequence homology, which seeks correlations between amino-acid sequences. Instead, correlations are sought for the inheritance of pairs of proteins into various species (for the phylogenetic profile method), for protein domains that exist both as fusions to each other and as free-standing polypeptides (for the Rosetta Stone method), or for the position of genes on chromosomes (for the gene neighbour method). Analysis of genomic and expression data by such methods produces networks of functional linkages between proteins in cells, and alters fundamentally the notion of what is meant by 'the function of a protein'.

Proteins are the main catalysts, structural elements, signalling messengers and molecular machines of biological tissues. Until recently, there have been two principal ways to learn more about the functions of protein molecules. All primary knowledge of function has come from some biochemical, genetic or structural experiment on an individual protein. But once a function has been assigned to an individual protein, one can search for other proteins with related functions by seeking proteins whose amino-acid sequences are similar to the original protein. This 'homology method' is used widely to extend knowledge of protein function from one protein to its cousins, which are presumably descended from the same common ancestral protein. The powerful BLAST programs¹ are used to extend experimental knowledge of protein function to new sequences in this way. By using such homology methods, roughly 40–70% of new genome sequences can be assigned to some function, the larger percentage being for well-studied prokaryotes^{2–4}. The functional assignments by homology usually involve identification of some molecular function of the protein, but they do not place the protein in its context of cellular function, as do the methods described below.

New methods have been devised to supply functional information for many proteins at once. In some cases, assignments can be made to most of the proteins encoded by the genome of an organism. These methods often detect a functional linkage between proteins. If the function of one of the proteins is known, then it can be inferred that the

Box 1

The method of phylogenetic profiles



The method of phylogenetic profiles is illustrated with four hypothetical genomes (top), each containing a subset of several proteins labelled P1, ..., P7. The presence or absence of each protein is indicated by 1 or 0, respectively, in the phylogenetic profiles given on the lower left. Identical profiles are clustered in boxes on the right, with profiles differing by one bit connected by lines. The conclusion at the bottom is that proteins P2 and P7 are functionally linked because they have the same phylogenetic profile and, similarly, that proteins P3 and P6 are functionally linked. Notice that two proteins that are functionally linked in this way are not in general homologues: they do not require similar sequences. This method has been described in ref. 16 with related concepts given in refs 21–25.

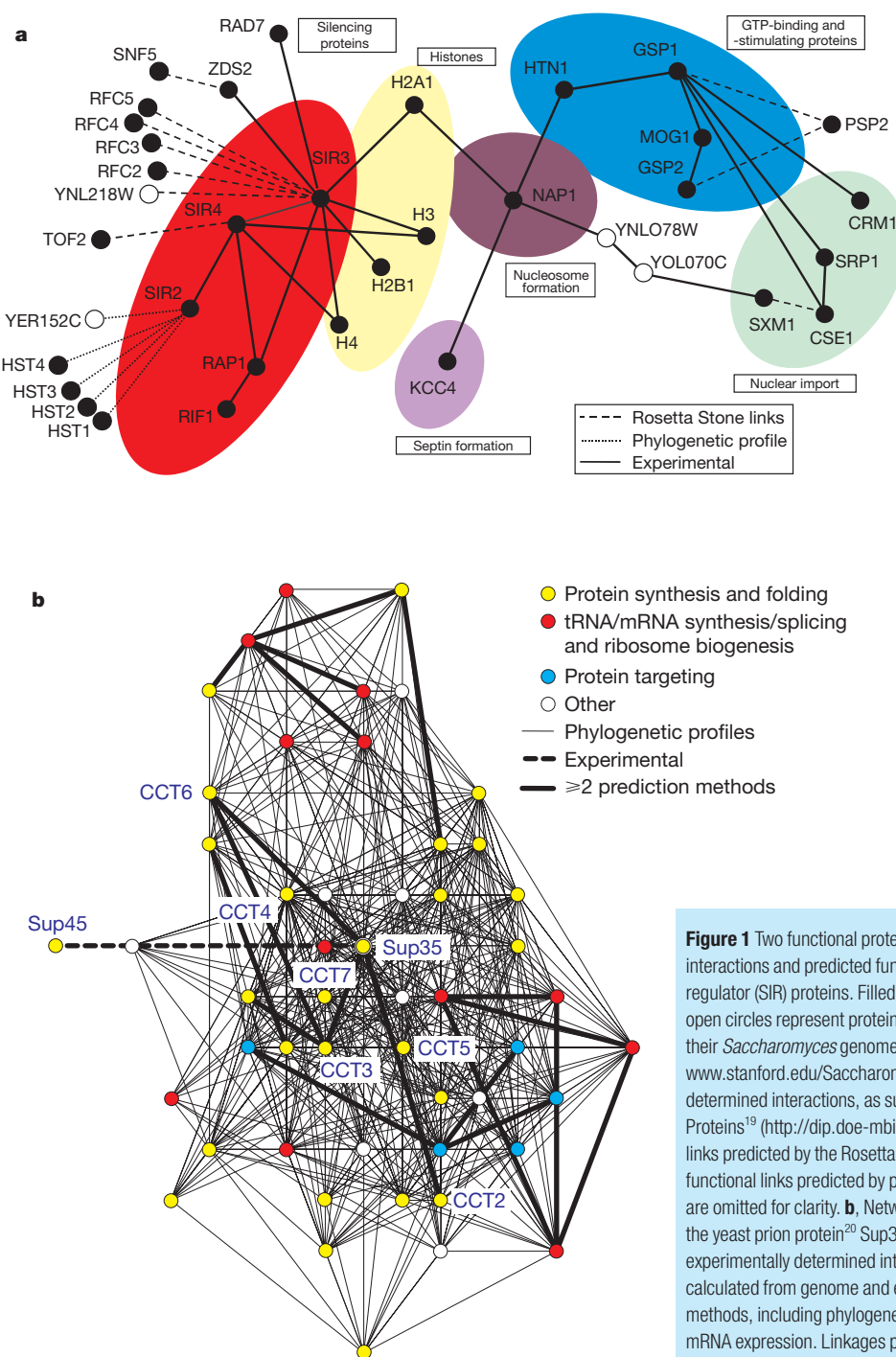


Figure 1 Two functional protein networks. **a**, Network of protein interactions and predicted functional links involving silencing information regulator (SIR) proteins. Filled circles represent proteins of known function; open circles represent proteins of unknown function, represented only by their *Saccharomyces* genome sequence numbers (<http://genome-www.stanford.edu/Saccharomyces>). Solid lines show experimentally determined interactions, as summarized in the Database of Interacting Proteins¹⁹ (<http://dip.doe-mbi.ucla.edu>). Dashed lines show functional links predicted by the Rosetta Stone method¹². Dotted lines show functional links predicted by phylogenetic profiles¹⁶. Some predicted links are omitted for clarity. **b**, Network of predicted functional linkages involving the yeast prion protein²⁰ Sup35. The dashed line shows the only experimentally determined interaction. The other functional links were calculated from genome and expression data¹¹ by a combination of methods, including phylogenetic profiles, Rosetta stone linkages and mRNA expression. Linkages predicted by more than one method, and hence particularly reliable, are shown by heavy lines. Adapted from ref. 11.

linked proteins act in the same pathway or complex as the first protein. Even if none of the linked proteins has a known function, knowledge of the linkages is valuable in focusing future experiments and adding to the infrastructure of cellular function.

One of the most powerful of the new methods extends the two-hybrid screen to a genome-wide assay and has detected over 1,000 putative protein–protein interactions in yeast cells (see review in this issue by Pandey and Mann, pp. 837–846, and refs 5, 6). Another powerful class of methods is the analysis of correlated mRNA expression levels (see review by Lockhart and Winzler, pp. 827–836, and refs 7–9). These methods detect changes in mRNA expression in different cell types, such as a B-cell lymphoma compared with normal cells, or in

yeast cells challenged by metabolic or environmental conditions (for instance, starvation or heat). By correlating those mRNAs whose expression levels are changed, one can establish functional linkages between the proteins encoded by the correlated mRNAs^{10,11}.

Computational detection of functional linkages

The advent of fully sequenced genomes has facilitated the development of computational methods for establishing functional linkages between proteins. One of these computational methods is the phylogenetic profile (Box 1). A phylogenetic profile describes the pattern of presence or absence of a particular protein across a set of organisms whose genomes have been sequenced. If two proteins have

Box 2

The Rosetta Stone method for detecting functional linkage

General concept


C. elegans

E. coli TrpC


The domain fusion or Rosetta Stone method for detecting functional linkage^{12,15} is illustrated here by three examples. The top sequence in all three triplets of proteins is the fused domain or Rosetta Stone sequence; it is homologous to two separate sequences in another species. In the middle example, the genes *Pur2* and *Pur3* of yeast both encode enzymes that catalyse steps in the purine biosynthetic pathway. If it were not previously known from biochemical and genetic experiments that these enzymes are functionally linked, the linkage would be apparent from the Rosetta stone sequence Ade5,7,8 from *Caenorhabditis elegans*. Similarly, in the lower example, the fused sequence of TrpC in the *Escherichia coli* genome would inform us that the yeast proteins TrpG and TrpF are functionally linked, if we did not know already that they both catalyse steps in the biosynthesis of tryptophan.

the same phylogenetic profile (that is, the same pattern of presence or absence) in all surveyed genomes, it is inferred that the two proteins have a functional link. That is, why would two proteins always both be inherited into a new species, or neither inherited, unless the two function together? The power of the method to detect functional linkage can be appreciated when the number of possible phylogenetic profiles is considered: because each protein can be either present or absent in each genome, if there are n fully sequenced genomes, there are up to 2^n phylogenetic profiles. Currently there are about 30 fully sequenced genomes in the public domain, meaning there are 2^{30} ($\sim 10^9$) possible phylogenetic profiles. This number far exceeds the number of protein families, so that a protein's phylogenetic profile is a nearly unique characterization of its pattern of distribution among genomes. Hence any two proteins having identical or similar phylogenetic profiles are likely to be engaged in a common pathway or complex.

Functional linkages between proteins have also been detected by analysing fusion patterns of protein domains (Box 2). Not infrequently, separate proteins A and B in one organism are expressed as a fused protein in some other species. When expressed as a fused protein, the two domains A and B are almost certainly linked in function. Thus a successful search through other genome sequences for the corresponding fused protein is powerful evidence that A and B are linked functionally. Because A and B have unrelated sequences, this type of functional linkage cannot be detected by a homology search. Also, because the fused protein has similarity to both A and B, it is termed a Rosetta Stone sequence¹².

A third computational method that reveals functional linkages from genome sequences is the gene neighbour method^{13,14}. If in several genomes the genes that encode two proteins are neighbours on the chromosome, the proteins tend to be functionally linked. This method can be powerful in uncovering functional linkages in prokaryotes, where operons are common, but also shows promise for analysing interacting proteins in eukaryotes (Box 3).

Functional networks

When methods for detecting functional linkages are applied to all the proteins of an organism^{11,15}, networks of interacting, functionally linked proteins can be traced out. Two examples from yeast are given in Fig. 1. Figure 1a shows interactions among histones and related proteins such as silencing proteins. These were determined mostly by experiments, but some links were predicted by the Rosetta Stone method and by phylogenetic profiles. Some of the links are to proteins known only from their genome sequences, and without other functional information; their linkage to this network indicates an intimate functional interaction among proteins involved in gene silencing, DNA packaging and nuclear transport.

Figure 1b shows a second network of functionally linked proteins

from yeast, centred on the yeast prion protein Sup35. In this network, most of the links are predicted by phylogenetic profiles, the Rosetta stone method and mRNA expression patterns. Sup35 is known to regulate translation, and it is therefore of interest that most of the predicted linkages are to other proteins involved in protein synthesis, folding and targeting. This indicates that at least some of the predicted links are meaningful. As methods improve for detecting protein linkages, it seems likely that most yeast proteins will be included in expanded versions of the networks of Fig. 1. A central feature of these networks is that most proteins interact with several other proteins.

Validation of functional linkages

What evidence is there that functional linkages predicted by phylogenetic profiles, Rosetta stone and related methods are valid? At first glance, there is the reassurance that these methods link many proteins that are already known to function together on the basis of experiments. Examples include ribosomal proteins, proteins from the flagellar motor apparatus, and proteins in known metabolic pathways^{11,16}. A more quantitative validation is offered by the check of 'keyword recovery'¹¹. This simple assay compares the keyword annotations¹⁷ for both members of each pair of proteins linked by one

Box 3

The method of correlated gene neighbours for inferring functional linkage

Observed gene locations



Inferred functional linkage



If two genes (blue and yellow in the figure) are found to be neighbours in several different genomes, a functional linkage may be inferred between the proteins they encode. The method is most robust for microbial genomes but may work to some extent even for human genes where operon-like clusters are observed (see, for example, ref. 26). The gene neighbour method correctly identifies functional links among eight enzymes in the biosynthetic pathway for arginine in *Mycobacterium tuberculosis*.

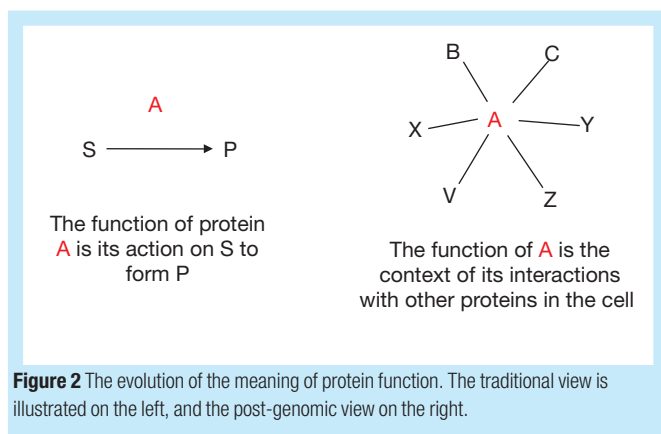


Figure 2 The evolution of the meaning of protein function. The traditional view is illustrated on the left, and the post-genomic view on the right.

of the methods. This is possible in those cases where both members of the pair have known functions. When the keywords for both members agree, there is said to be 'keyword recovery'. When keyword recovery was examined for the predicted functional linkages between yeast proteins, it was found that the individual methods showed an average signal-to-noise ratio for keyword recovery ranging between 2, for correlated mRNA expression, to 5, for the phylogenetic profiles. These values can be compared with that of 8 for direct experimental measurements of linkage. It was also found that when two of the predictive methods gave the same linkage, the signal-to-noise value was 8, the same as for direct experiments. In short, the computer-based methods for inferring function have fair reliability in general, and excellent reliability when two or more of them agree on a link.

The post-genomic view of function

The classical view of protein function focuses on the action of a single protein molecule. This action may be the catalysis of a given reaction or the binding of a small or large molecule. Today this local function is sometimes termed the 'molecular function' of the protein to distinguish it from an expanded view of function (Fig. 2). In the expanded view of protein function, a protein is defined as an element in the network of its interactions. Various terms have been coined for this expanded notion of function, such as 'contextual function' or 'cellular function' (see, for example, ref. 18). Whatever the term, the idea is that each protein in living matter functions as part of an extended web of interacting molecules.

In conclusion, the availability of fully sequenced genomes and the enormous amount of data on the co-expression of mRNAs opens new ways to analyse protein function. The new methods establish functional links between pairs of proteins, and interconnecting links form networks of functionally interacting proteins. Some of the functional

linkages reflect metabolic or signalling pathways; other linkages reflect the formation of complexes of macromolecules such as ribosomes. Often it is possible to understand the cellular functions of uncharacterized proteins through their linkages to characterized proteins. In broader terms, the networks of linkages offer a new view of the meaning of protein function, and in time should offer a deepened understanding of the functioning of cells. □

1. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
2. The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
3. Chervitz, S. A. *et al.* Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* **282**, 2022–2028 (1998).
4. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
5. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
6. Ito, T. *et al.* Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA* **97**, 1143–1147 (2000).
7. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
8. Lashkari, D. A. *et al.* Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl Acad. Sci. USA* **94**, 13057–13062 (1997).
9. Brown, P. O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21**, 33–37 (1999).
10. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
11. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).
12. Marcotte, E. M. *et al.* Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
13. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328 (1998).
14. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA* **96**, 2896–2901 (1999).
15. Enright, A. J., Iliopoulos, I., Kyripides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86–90 (1999).
16. Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96**, 4285–4288 (1999).
17. Andrade, M. A. & Valencia, A. Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics* **14**, 600–607 (1998).
18. Kim, S. H. Structural genomics of microbes: an objective. *Curr. Opin. Struct. Biol.* (in the press).
19. Xenarios, I. *et al.* DIP: the Database of Interacting Proteins. *Nucleic Acids Res.* **28**, 289–291 (2000).
20. Wickner, R. B. [URE3] as an altered URE2 protein: evidence for a prion analog in *Saccharomyces cerevisiae*. *Science* **264**, 566–569 (1994).
21. Bork, P. *et al.* Predicting function: from genes to genomes and back. *J. Mol. Biol.* **283**, 707–725 (1998).
22. Huynen, M., Dandekar, T. & Bork, P. Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Lett.* **426**, 1–5 (1998).
23. Hegyi, H. & Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164 (1999).
24. Ouzounis, C. & Kyripides, N. The emergence of major cellular processes in evolution. *FEBS Lett.* **390**, 119–123 (1996).
25. Gaasterland, T. & Ragan, M. A. Constructing multigenome views of whole microbial genomes. *Microb. Comp. Genomics* **3**, 177–192 (1998).
26. Wu, Q. & Maniatis, T. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* **97**, 779–790 (1999).

Genomics, gene expression and DNA arrays

David J. Lockhart & Elizabeth A. Winzeler

Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, California 92121, USA

Experimental genomics in combination with the growing body of sequence information promise to revolutionize the way cells and cellular processes are studied. Information on genomic sequence can be used experimentally with high-density DNA arrays that allow complex mixtures of RNA and DNA to be interrogated in a parallel and quantitative fashion. DNA arrays can be used for many different purposes, most prominently to measure levels of gene expression (messenger RNA abundance) for tens of thousands of genes simultaneously. Measurements of gene expression and other applications of arrays embody much of what is implied by the term 'genomics'; they are broad in scope, large in scale, and take advantage of all available sequence information for experimental design and data interpretation in pursuit of biological understanding.

Biological and biomedical research is in the midst of a significant transition that is being driven by two primary factors: the massive increase in the amount of DNA sequence information and the development of technologies to exploit its use. Consequently, we find ourselves at a time when new types of experiments are possible, and observations, analyses and discoveries are being made on an unprecedented scale. Over the past few years, more than 30 organisms have had their genomes completely sequenced, with another 100 or so in progress (see www.tigr.org or genomes@ncbi.nlm.nih.gov for a list). At least partial sequence has been obtained for tens of thousands of mouse, rat and human genes, and the sequence of two entire human chromosomes (chromosomes 21 and 22) has been determined^{1,2}. Within the year, a large proportion of the human genome will be deciphered, in both public and private efforts, and the complete sequence of the mouse and other animal and plant genomes will undoubtedly follow close behind. Unfortunately, the billions of bases of DNA sequence do not tell us what all the genes do, how cells work, how cells form organisms, what goes wrong in disease, how we age or how to develop a drug. This is where functional genomics comes into play. The purpose of genomics is to understand biology, not simply to identify the component parts, and the experimental and computational methods take advantage of as much sequence information as possible. In this sense, functional genomics is less a specific project or programme than it is a mindset and general approach to problems. The goal is not simply to provide a catalogue of all the genes and information about their functions, but to understand how the components work together to comprise functioning cells and organisms.

To take full advantage of the large and rapidly increasing body of sequence information, new technologies are required. Among the most powerful and versatile tools for genomics are high-density arrays of oligonucleotides or complementary DNAs. Nucleic acid arrays work by hybridization of labelled RNA or DNA in solution to DNA molecules attached at specific locations on a surface. The hybridization of a sample to an array is, in effect, a highly parallel search by each molecule for a matching partner on an 'affinity matrix',

with the eventual pairings of molecules on the surface determined by the rules of molecular recognition. Arrays of nucleic acids have been used for biological experiments for many years³⁻⁸. Traditionally, the arrays consisted of fragments of DNA, often with unknown sequence, spotted on a porous membrane (usually nylon). The arrayed DNA fragments often came from cDNA, genomic DNA or plasmid libraries, and the hybridized material was often labelled with a radioactive group. Recently, the use of glass as a substrate and fluorescence for detection, together with the development of new technologies for synthesizing or depositing nucleic acids on glass slides at very high densities, have allowed the miniaturization of nucleic acid arrays with concomitant increases in experimental efficiency and information content⁹⁻¹⁴ (Fig. 1).

While making arrays with more than several hundred elements was until recently a significant technical achievement, arrays with more than 250,000 different oligonucleotide probes or 10,000 different cDNAs per square centimetre can now be produced in significant numbers^{15,16}. Although it is possible to synthesize or deposit DNA fragments of unknown sequence, the most common implementation is to design arrays based on specific sequence information, a process sometimes referred to as 'downloading the genome onto a chip' (Fig. 1). There are several variations on this basic technical theme: the hybridization reaction may be driven (for example, by an electric field)^{17,18}; other detection methods¹⁹ besides fluorescence can be used; and the surface may be made of materials other than glass such as plastic, silicon, gold, a gel or membrane, or may even be comprised of beads at the ends of fibre-optic bundles²⁰⁻²². Nonetheless, the key elements of parallel hybridization to localized, surface-bound nucleic acid probes and subsequent counting of bound molecules are ubiquitous, and high-density arrays of nucleic acids on glass (often called DNA microarrays, oligonucleotide arrays, GeneChip arrays, or simply 'chips') and their biological uses will be the focus of this review.

Global gene expression experiments

One of the most important applications for arrays so far is the monitoring of gene expression (mRNA abundance). The collection of genes that are expressed or transcribed from genomic DNA, sometimes referred to as the expression

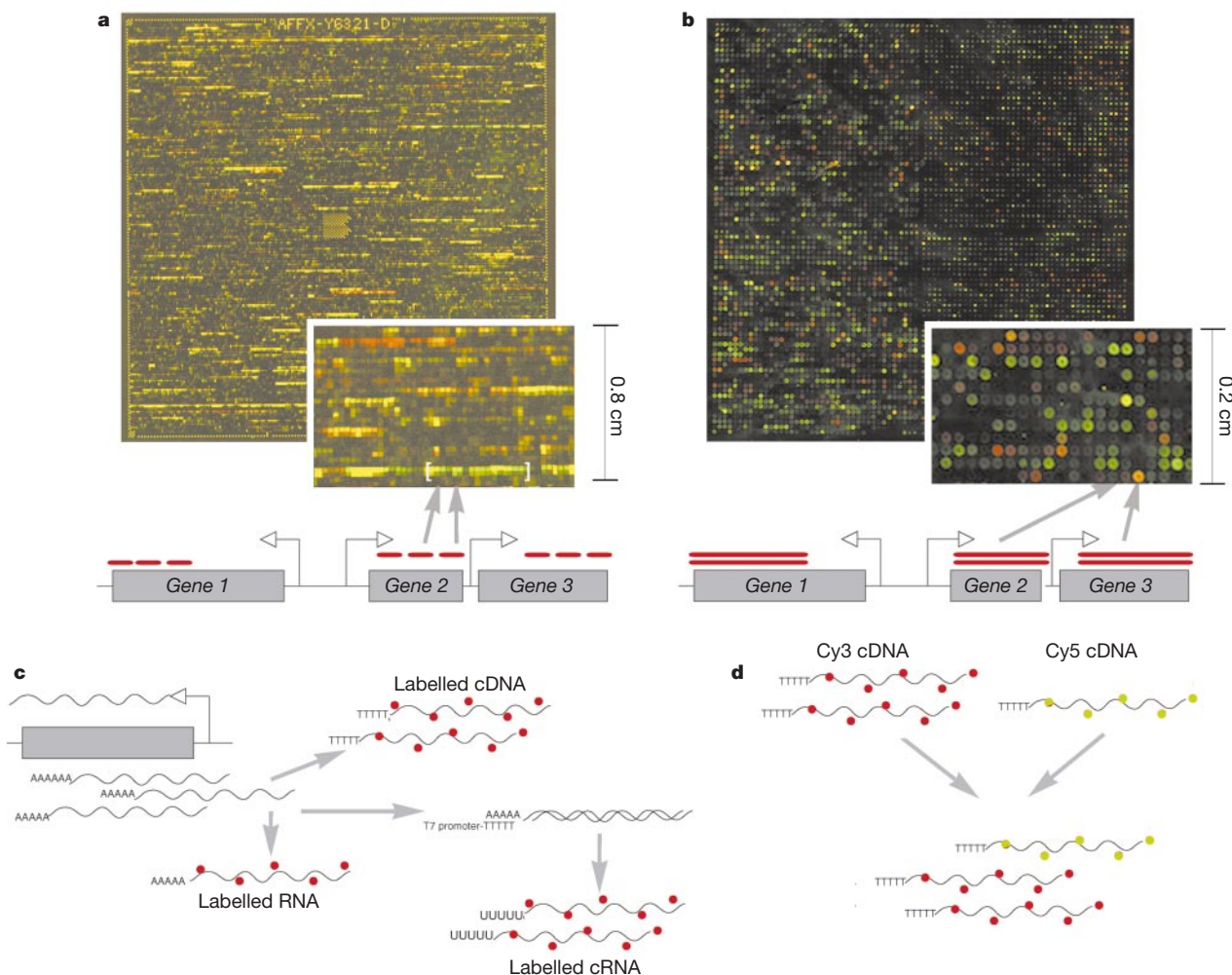
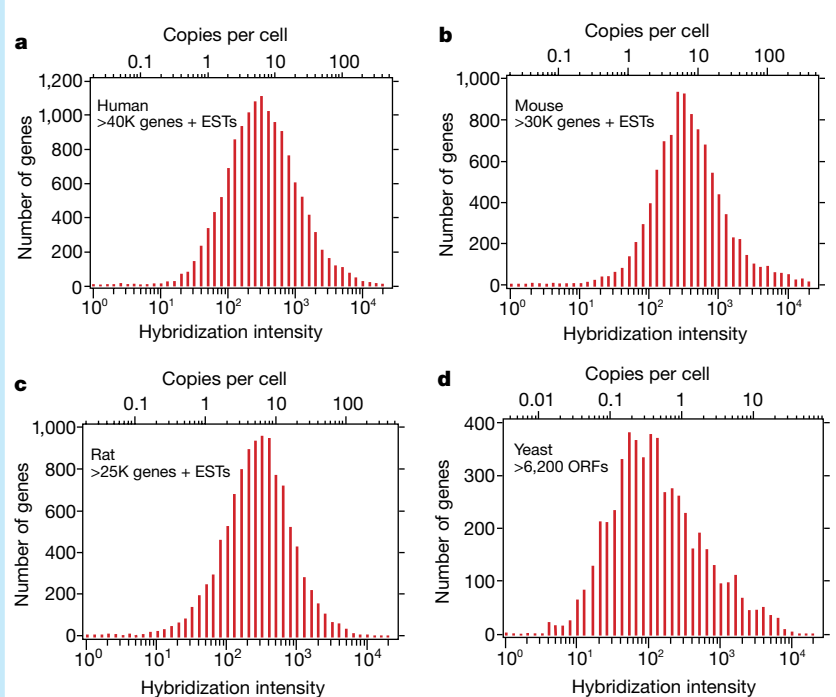


Figure 1 Principal types of arrays used in gene expression monitoring. Nucleic acid arrays are generally produced in one of two ways: by robotic deposition of nucleic acids (PCR products, plasmids or oligonucleotides) onto a glass slide²⁵ or *in situ* synthesis (using photolithography¹⁵) of oligonucleotides. Shown are pseudocolour images of **a**, an oligonucleotide array and **b**, a cDNA array after hybridization of labelled samples and fluorescence detection. In both cases the images have been coloured to indicate the relative number of yeast transcripts present under two different growth conditions (red, high in condition 1, low in condition 2; green, high in condition 2, low in condition 1; yellow, high under both conditions; black, low under both conditions). In the case of photolithographically synthesized arrays, $\sim 10^7$ copies of each selected oligonucleotide (usually 20 to 25 nucleotides in length) are synthesized base by base in hundreds of thousands of different $24 \mu\text{m} \times 24 \mu\text{m}$ areas on a $1.28 \text{ cm} \times 1.28 \text{ cm}$ glass surface. For robotic deposition, approximately one nanogram of material is deposited at intervals of 100–300 μm . Typically for oligonucleotide arrays, multiple probes per gene are placed on the array (20 pairs in the example shown here), while in the case of robotic deposition, a single, longer (up to 1,000 bp) double-stranded DNA probe is used for each gene or EST. In both cases, probes are usually designed from sequence located nearer to the 3' end of the gene (near the poly-A tail in eukaryotic mRNA), and different probes can be used for different exons. After hybridization of labelled samples (typically overnight), the arrays are scanned and the quantitative fluorescence image along with the known identity of the probes is used to assess the 'presence' or 'absence' (more precisely, the detectability above thresholds based on background and noise levels) of a particular molecule (such as a transcript), and its relative abundance in one or more samples. Because the sequence of the oligonucleotide or cDNA at each physical location (or address) is generally known or can be determined, and because

the recognition rules that govern hybridization are well understood, the signal intensity at each position gives not only a measure of the number of molecules bound, but also the likely identity of the molecules. Although oligonucleotide probes vary systematically in their hybridization efficiency, quantitative estimates of the number of transcripts per cell can be obtained directly by averaging the signal from multiple probes^{15,26,30}. For technical reasons, the information obtained from spotted cDNA arrays gives the relative concentration (ratio) of a given transcript in two different samples (derived from competitive, two-colour hybridizations). Messenger RNAs present at a few copies (relative abundance of $\sim 1:100,000$ or less) to thousands of copies per mammalian cell can be detected^{25,26,30}, and changes as subtle as a factor of 1.3 to 2 can be reliably detected if replicate experiments are performed. **c**, Different methods for preparing labelled material for measurements of gene expression. The RNA can be labelled directly, using a psoralen–biotin derivative or by ligation to an RNA molecule carrying biotin²⁶; labelled nucleotides can be incorporated into cDNA during or after reverse transcription of polyadenylated RNA; or cDNA can be generated that carries a T7 promoter at its 5' end. In the last case, the double-stranded cDNA serves as template for a reverse transcription reaction in which labelled nucleotides are incorporated into cRNA. Commonly used labels include the fluorophores fluorescein, Cy3 (or Cy5), or nonfluorescent biotin, which is subsequently labelled by staining with a fluorescent streptavidin conjugate. **d**, Two-colour hybridization strategy often used with cDNA microarrays. cDNA from two different conditions is labelled with two different fluorescent dyes (usually Cy3 and Cy5), and the two samples are co-hybridized to an array. After washing, the array is scanned at two different wavelengths to detect the relative transcript abundance for each condition. cDNA array image courtesy of J. DeRisi and P. O. Brown (<http://cmgm.stanford.edu/pbrown/yeastchip.html>).

Figure 2 Messenger RNA abundance levels in different cells, tissues and organisms. **a**, Human HIV-infected T lymphocytes; **b**, mouse olfactory epithelium; **c**, rat brain; **d**, *S. cerevisiae* strain RY136 grown at 25 °C in rich medium. Levels of gene expression were measured using Affymetrix oligonucleotide arrays. For human, mouse and rat samples, hybridization intensities were converted to copies per cell (top axis) based on the signal from multiple control RNAs added to the samples at known concentrations. For yeast, the conversion was based on the signal from the TATA-binding protein (TBP) mRNA, which has been determined to be present at ~3.5 copies per cell when yeast cells are grown in rich medium¹⁰³. Only those genes scored as 'present' are represented in the histograms. Data from multiple arrays containing probes for a different subset of genes and ESTs were combined to generate the plots for human (five arrays), mouse (five arrays) and rat (three arrays). All yeast ORFs were represented on a single array. For measurements that cover such a large number of genes, it is important to maintain high standards of data quality to keep false-positive results to a minimum. (For example, when monitoring 10,000 genes, even a low false-positive rate of 1% results in 100 false calls.) We find that the source of most false positives (in large part the result of setting the lowest possible thresholds in the interest of sensitivity) is random noise, biological variation, or the occasional array-specific physical defect, so observations made consistently in independent replicates yield a

false-positive rate close to 0.01%, or only 1 in 10,000. In well controlled experiments involving specific biochemical, chemical and genetic perturbations, typically the number of expression differences is modest, with about 0.1–2% of the monitored genes changing by a factor of 1.8 or more, and only a small fraction of these changing by more than four- to fivefold^{56–58,70–72,95,104}. For samples derived, for example, from different adult human or mouse tissues, or from normal versus advanced tumour tissue, the number of differences can be as large as 10–15% of the monitored genes^{50–53}. The larger number of differences poses only minor difficulties for the technology, but analysis of the more complex results and the larger number of genes involved typically requires more sophisticated computational methods.



profile or the 'transcriptome', is a major determinant of cellular phenotype and function. The transcription of genomic DNA to produce mRNA is the first step in the process of protein synthesis, and differences in gene expression are responsible for both morphological and phenotypic differences as well as indicative of cellular responses to environmental stimuli and perturbations. Unlike the genome, the transcriptome is highly dynamic and changes rapidly and dramatically in response to perturbations or even during normal cellular events such as DNA replication and cell division^{23,24}. In terms of understanding the function of genes, knowing when, where and to what extent a gene is expressed is central to understanding the activity and biological roles of its encoded protein. In addition, changes in the multi-gene patterns of expression can provide clues about regulatory mechanisms and broader cellular functions and biochemical pathways. In the context of human health and treatment, the knowledge gained from these types of measurements can help determine the causes and consequences of disease, how drugs and drug candidates work in cells and organisms, and what gene products might have therapeutic uses themselves or may be appropriate targets for therapeutic intervention.

Past discussions of arrays have often centred on technical issues and specific performance characteristics²⁵. Now that nucleic acid arrays have been constructed for many different organisms^{14,26–29} and used successfully to measure transcript abundance in a host of different experiments, the focus of interest has thankfully shifted. Investigators are now more concerned with questions concerning experimental design, data analysis, the use of small amounts of mRNA from limited sources, the best ways to extract biological meaning from the results, pathway and cell-circuitry modelling, and medical uses of expression patterns.

Array-based gene expression monitoring

One way to think of measurements with arrays is that they are simply a more powerful substitute for conventional methods of evaluating

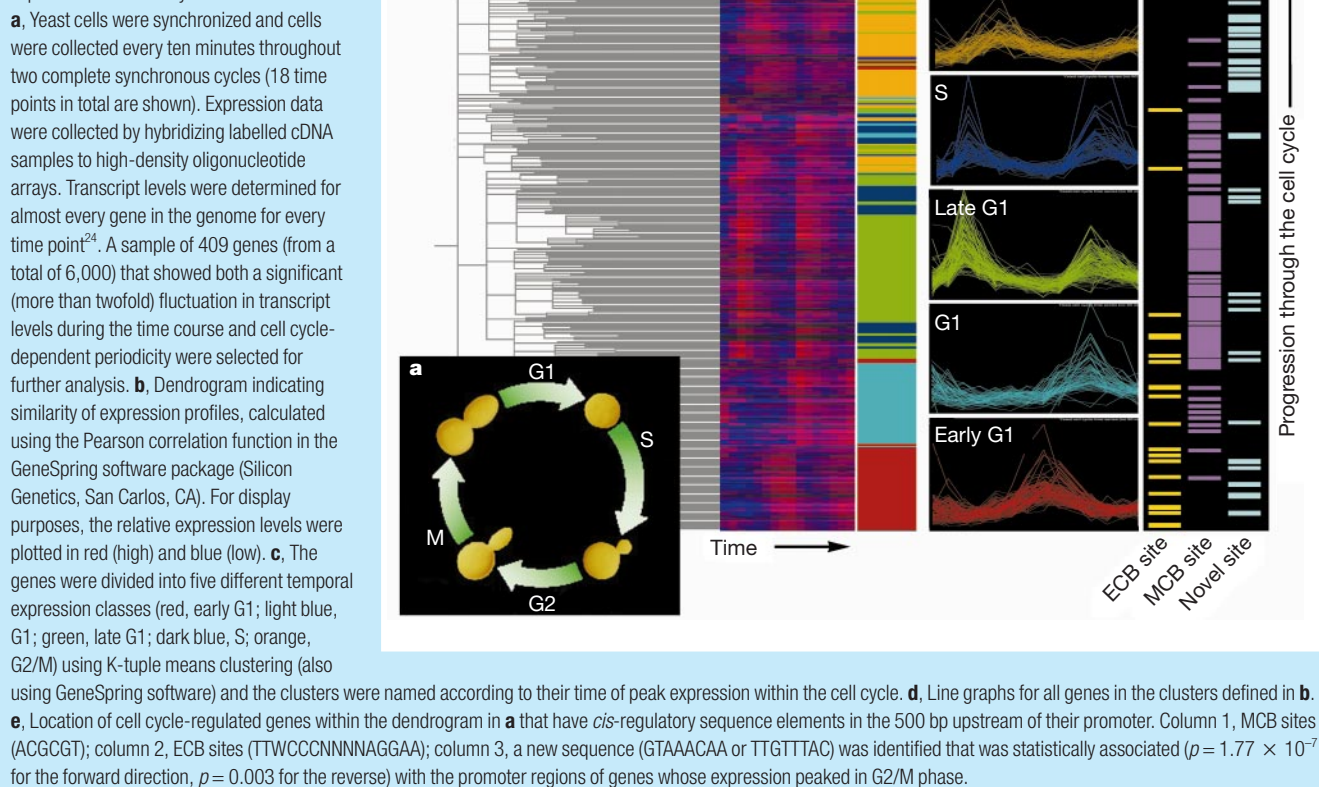
mRNA abundance. For some early experiments, only a relatively small set of genes, which were thought to be important to a process, were included on the arrays^{12,30}. However, such experiments did not capitalize on the arrays' potential: a key advantage of using arrays, especially those that contain probes for tens of thousands of different genes, is that it is not necessary to guess what the important genes or mechanisms are in advance. Instead of looking only under the proverbial lamppost, a broader, more complete and less biased view of the cellular response is obtained (Figs 2, 3).

The breadth of array-based observations almost guarantees that surprising findings will be made. A recent study measured the transcriptional changes that occur as cells progress through the normal cell-division cycle in humans for approximately 40,000 genes (R. J. Cho *et al.*, unpublished results). In addition to the induction of DNA replication genes and genes involved with cell-cycle control and chromosome segregation that would be expected at specific stages in the cell cycle, a large collection of genes involved with smooth muscle function, apoptosis and intercellular adhesion and cell motility were found to be upregulated during a specific phase. The expected results act effectively as internal controls that provide a certain amount of validation (and comfort), while new information is obtained by a systematic search of a larger part of 'gene space'. In addition, because arrays often contain probes for genes of unknown function (and often with only partial sequence information), any outcome for these could be considered, in some sense, both surprising and novel (although clearly requiring further characterization).

Other gene expression methods

Not surprisingly, there are other ways to measure mRNA abundance, gene expression and changes in gene expression. For measuring gene expression at the level of mRNA, northern blots, polymerase chain reaction after reverse transcription of RNA (RT-PCR), nuclease

Figure 3 Methods for analysing gene expression data shown for measurements of expression in the cell cycle of *S. cerevisiae*.



protection, cDNA sequencing, clone hybridization, differential display³¹, subtractive hybridization, cDNA fragment fingerprinting^{32–35} and serial analysis of gene expression (SAGE)³⁶ have all been put to good use to measure the expression levels of specific genes, characterize global expression profiles or to screen for significant differences in mRNA abundance. But if messenger RNA is only an intermediate on the way to production of the functional protein products, why measure mRNA at all? One reason is simply that protein-based approaches are generally more difficult, less sensitive and have a lower throughput than RNA-based ones. But more importantly, mRNA levels are immensely informative about cell state and the activity of genes, and for most genes, changes in mRNA abundance are related to changes in protein abundance. Because of its importance, however, many methods have been developed for monitoring protein levels either directly or indirectly (see review in this issue by Pandey and Mann, pages 837–846). These include western blots, two-dimensional gels, methods based on protein or peptide chromatographic separation and mass spectrometric detection^{37–40}, methods that use specific protein-fusion reporter constructs and colorimetric readouts^{41–44}, and methods based on characterization of actively translated, polysomal mRNA^{45–47}.

The importance of the protein-based methods is that they measure the final expression product rather than an intermediate. In addition, some of them enable the detection of post-translational protein modifications (for example, phosphorylation and glycosylation) and protein complexes, and in some cases, yield information about protein localization, none of which are obtained directly by measurements of mRNA. There is no question that protein- and RNA-based measurements are complementary, and that protein-based methods are important as they measure observables that are not readily detected in other ways.

Human disease, gene expression and discovery

Genomics and gene expression experiments are sometimes derided as ‘fishing expeditions’. Our view is that there is nothing wrong with a

fishing expedition⁴⁸ if what you are after is ‘fish’, such as new genes involved in a pathway, potential drug targets or expression markers that can be used in a predictive or diagnostic fashion. Because the arrays can be designed and made on the basis of only partial sequence information, it is possible to include genes in a survey that are completely uncharacterized. In many ways, the spirit of this approach is more akin to that of classical genetics in which mutations are made broadly and at random (not only in specific genes), and screens or selections are set up to discover mutants with an interesting phenotype, which then leads to further characterization of specific genes.

Such broad discovery experiments are probably better described as ‘question-driven’ rather than hypothesis-driven in the conventional sense. But that is not to diminish their value for understanding basic biological processes and even for understanding and treating human disease. For example, by analysing multiple samples obtained from individuals with and without acute leukaemia or diffuse large B-cell lymphoma, gene expression (mRNA) markers were discovered that could be used in the classification of these cancers^{49,50}. The importance of monitoring a large number of genes was well illustrated in these studies. Golub *et al.*⁴⁹ found that reliable predictions could not be made based on any single gene, but that predictions based on the expression levels of 50 genes (selected from the more than 6,000 monitored on the arrays) were highly accurate. The results of both of these studies indicate that measurements with more individuals and more genes will be needed to identify robust expression markers that are predictive of clinical outcome. But even with the limited initial data it was possible to help clarify an unusual case (classic leukaemia presentation but atypical morphology) and to use this information to guide the patient’s clinical care.

It is also possible to take a related approach to help understand what goes wrong in cancerous, transformed cells and to identify the genes responsible for disease. Causative effects and potential

therapeutic targets can be identified by determining which genes are upregulated in different tumour types^{51–55}, and specific candidate genes can be intentionally overexpressed in cell lines or cells treated with growth factors in order to identify downstream target genes and to explore signalling pathways^{56–58}. Tumorigenesis is often accompanied by changes in chromosomal DNA, such as genetic rearrangements, amplifications or losses of particular chromosomal loci, and developmental abnormalities, such as Down's or Turner's syndrome, may arise from aberrations in DNA copy number. Because genomic DNA can be interrogated in much the same way as mRNA, comparisons of the copy number of genomic regions or the genotype of genetic markers can be used to detect chromosomal regions and genes that are amplified or deleted in cancerous or pre-cancerous cells. By using arrays containing probes for a large number of genes or polymorphic markers, changes in DNA copy number have been detected in both breast cancer cell lines and in tumours^{59–61}. The identification of when and where changes in copy number or chromosomal rearrangements have occurred can be used in both the classification of cancer types and the identification of regions that may harbour tumour-suppressor genes.

Whole-genome hypotheses

The use of genomics tools such as arrays does not, of course, preclude hypothesis-driven research. For fully sequenced organisms, arrays containing probes for every annotated gene in the genome have been produced^{14,26}. With these one can ask, for example, whether a transcription factor has a global role in transcription (affecting all genes) or a specific role (affecting only some). Holstege *et al.*⁶² used this type of application in a genome-wide expression analysis in yeast to functionally dissect the machinery of transcription initiation. Similarly, genes located near the ends of chromosomes in yeast (as well as genes at the mating-type locus) are known to be transcriptionally 'silent'. Full genome arrays allow the chromosomal landscape of silencing to be mapped, and make it possible to test whether what is true for a handful of well-studied genes near the telomeres is true for all telomeric genes, and whether any centromere-proximal genes are also transcriptionally silenced⁶³.

It is important to emphasize that these new, parallel approaches do not replace conventional methods. Standard methods such as northern blots, western blots or RT-PCR are simply used in a more targeted fashion to complement the broader measurements and to follow-up on the genes, pathways and mechanisms implicated by the array results. Because the incidence of false-positive results can be made sufficiently low (see Fig. 2), it is not necessary to independently confirm every change for the results to be valid and trustworthy, especially if conclusions are based on changes in sets of genes rather than individual genes. More detailed follow-up is recommended if a gene is being chosen, for example, as a drug target, as a candidate for population genetics studies, or as the target for the construction of a knockout mouse.

Does gene expression indicate function?

As additional, uncharacterized open reading frames (ORFs) are identified in different organisms by the various genome sequencing projects, researchers have begun to ask whether the expression pattern for a gene can be used to predict the functional role of its protein product. An increasingly common approach involves using the gene expression behaviour observed over multiple experiments to first cluster genes together into groups (see Fig. 3), either by manual examination of the data²⁴, or by using statistical methods such as self-organizing maps⁶⁴, K-tuple means clustering or hierarchical clustering^{23,65,66}. The basic assumption underlying this approach is that genes with similar expression behaviour (for example, increasing and decreasing together under similar circumstances) are likely to be related functionally. In this way, genes without previous functional assignments can be given tentative assignments or assigned a role in a biological process based on the known functions of genes in the same

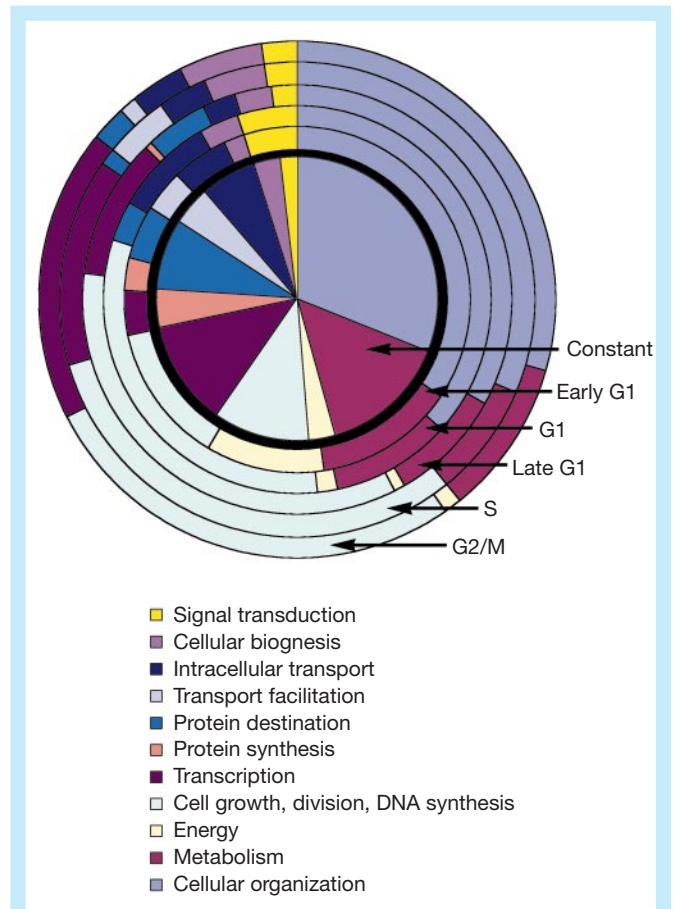


Figure 4 The 'guilt-by-association' method for assigning gene function. Functional distribution (using categories from MIPS: <http://www.mips.biochem.mpg.de/proj/yeast/catalogues/funcat/index.html>) of yeast genes whose periodic expression peaked at different times in the yeast cell cycle (outer rings) or was constant throughout the cell cycle (inner circle)²⁴. A much larger fraction of cell cycle-modulated genes is important in DNA synthesis, cell growth or cell division. Although there is a strong correlation between distinct expression profiles and functional assignments, specific expression behaviour should not be taken as sufficient evidence for functional assignment: not all genes involved in DNA replication are expressed periodically in the cell cycle, and some genes that do not need to be cell cycle-regulated are transcribed in a periodic fashion.

expression cluster (that is, the concept of 'guilt-by-association'). The validity of this approach has been demonstrated for many genes in *Saccharomyces cerevisiae*, a simple organism for which the entire genomic sequence and the functional roles of approximately 60% of the genes are known^{24,65,67} (Fig. 4). Although not logically rigorous, the utility of the guilt-by-association approach has been demonstrated, as genes already known to be related do, in fact, tend to cluster together based on their experimentally determined expression patterns (Fig. 4). The approach is made more systematic and statistically sound by calculating the probability that the observed functional distribution of differentially expressed genes could have happened by chance. The application of statistical rigour is essential to avoid overly subjective interpretations of the results based on the predispositions, prior knowledge and interests of the individual researcher.

A tentative functional assignment may not be much more than a low-resolution description or general classification. Descriptions of this type are similar to those that come out of more classical genetic screens and selections, which have provided the vast majority of functional annotations to date — they indicate that genes are involved with a particular cellular phenotype and that they are likely

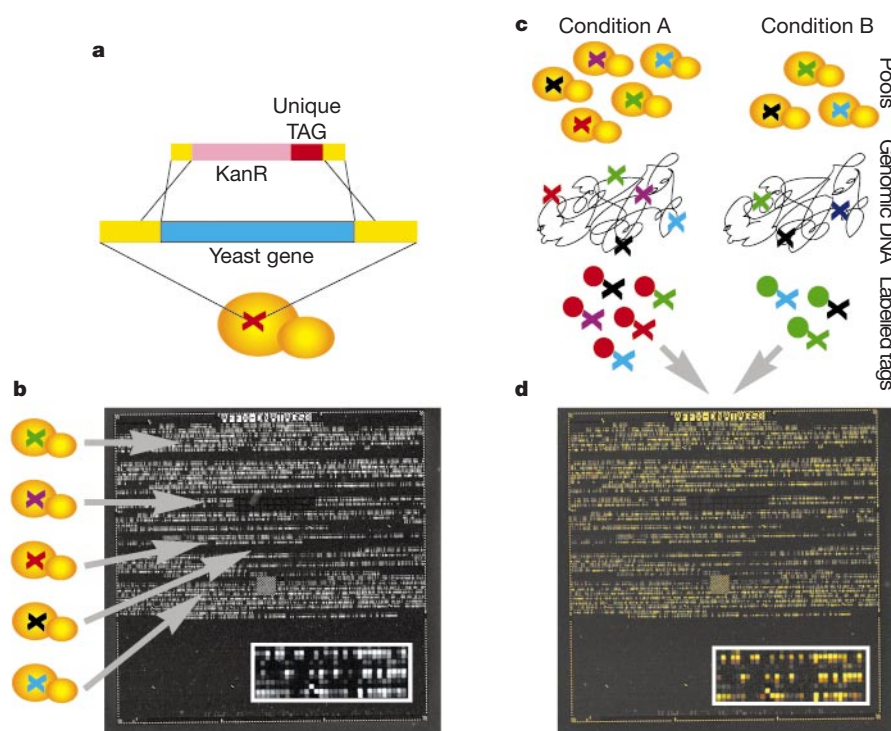


Figure 5 Generic oligonucleotide tag arrays for parallel phenotyping of mutant yeast strains. **a**, Many *S. cerevisiae* strains, each carrying a specific deletion of one of the more than 6,000 ORFs in the yeast genome, have been constructed⁹¹ by replacing individual genes with an antibiotic resistance cassette and a unique gene-specific 20-mer 'barcode', represented by an X. **b**, The barcode for each deletion strain corresponds to a specific location on an array that contains oligonucleotide probes that are complementary to the barcode sequences. **c**, Pools of different yeast strains can be assembled and grown under different conditions. After competitive growth, PCR is used to amplify the barcodes from genomic DNA isolated from the pools; the PCR products are subsequently labelled. **d**, By comparing the hybridization patterns of two different pools (before and after treatment with a drug, for example), the fitness of the strains can be assessed quantitatively. In this case, yeast genes required for sporulation or germination are represented in red, whereas yeast genes that are unnecessary for the process are shown in yellow. These same 20-mer sequences and the accompanying arrays are generic in design, and can be used to read the results of different types of 'bar-coded' reactions, such as those used for genotyping of human polymorphic loci¹⁰⁵. Images provided by R. M. Williams and R. W. Davis.

to be involved with a certain set of other genes and processes. This allows researchers to focus attention on a smaller subset of genes, many of which may not have been obvious candidates in the absence of the global expression observations. This overall approach highlights the importance of functional annotation and careful curation of existing sequence, function and knowledge databases (see below). Expression results covering thousands or even tens of thousands of genes and expressed sequence tags (ESTs) will be only partly interpretable given the functional and biological information available at the time they are initially generated. Our ability to extract knowledge from measurements of global gene expression tends to increase with time as additional information becomes available, and results can be subjected to further interrogation in the light of new information, observations, questions and hypotheses.

Gene expression and the regulation of transcription

When information on the complete genome sequence is available, as is the case for increasing numbers of small and even larger genomes, gene expression data can be used to identify new *cis*-regulatory elements (genomic sequence motifs that are over-represented in the genomic DNA in the vicinity of similarly behaving genes) and 'regulons' (sets of co-regulated genes), the basic units of the underlying cellular circuitry (Fig. 3d). In fact, the correlation between the presence of specific sequence motifs in promoter regions and gene expression patterns may be stronger than the correlation between functional categories and gene expression patterns. In yeast studies, more than 50% of the genes that are transcribed in a cell cycle-

specific manner and whose transcript abundance peaks in the G1 phase of the cell cycle have an MCB (Mlu cell-cycle box) within 500 base pairs (bp) of their translational start site^{24,68,69}. Similar observations have been made for yeast genes whose transcription is induced during sporulation⁶⁷. In addition, new *cis*-regulatory elements may be revealed by examining classes of co-regulated genes (Fig. 3d). With sufficiently large numbers of experimental observations of expression behaviour, the boundaries and all functioning sequence variants of *cis*-regulatory elements might be predicted without the need for the more conventional approach using site-directed mutagenesis ('promoter bashing'). The expression-based method will be especially valuable in exotic organisms, such as *Plasmodium falciparum*, the causative agent for malaria, for which experimental identification or verification of transcription factor binding sites is difficult.

Gene expression profiles as 'fingerprints'

An often overlooked aspect of measurements of global gene expression is that the sequence or even the origin of the arrayed probes does not need to be known to make interesting observations — the complex profiles, consisting of thousands of individual observations, can serve as transcriptional 'fingerprints'. The fingerprints can be used for classification purposes or as tests for relatedness, in a similar manner to the way in which DNA fingerprints are used in paternity testing. In one example, transcriptional fingerprints have been used to determine the target of a drug⁷⁰. The basic idea is that if a drug interacts with and inactivates a specific cellular protein, the phenotype of the drug-treated cell should be very similar to the phenotype

of a cell in which the gene encoding the protein has been genetically inactivated, usually through mutation. Thus, by comparing the expression profile of a drug-treated cell to the profiles of cells in which single genes have been individually inactivated, specific mutants can be matched to specific drugs, and therefore, targets to drugs. In a demonstration of this concept, the gene product of the *his3* gene was identified correctly as the target of 3-aminotriazole⁷⁰. Similarly, profiles have been used in the classification of cancers and the classification schemes did not depend on any specific information about the genes involved^{49,50}, although that information can be used to draw further biological and mechanistic conclusions. Finally, expression profiles can be used to classify drugs and their mode of action. For example, the functional similarity and specificity of different purine analogues have been determined by comparing the genome-wide effects on treated yeast, murine and human cells^{71,72}.

Expression measurements from small amounts of RNA

An important frontier in the development of gene expression technology involves reduction of the required amount of starting material. Most array-based expression measurements are done using RNA from a million or more cells, and obtaining such a relatively large sample is not a problem in many types of studies (for example, litres of yeast cells can be grown easily). However, in some cases, it is important or even necessary to use fewer cells, as when using a small organ from a fly or worm, sorted cells that express a rare marker, or laser-capture microdissected^{73–75} tumour tissue. Efficient and reproducible mRNA amplification methods are required, and there are two primary approaches that show significant promise. The first is a PCR-based approach that has been used to make single-cell cDNA libraries^{76–78}. We have found that the amplification is efficient and reproducible, but that the relative abundance of the cDNA products is not well correlated with the original mRNA levels (D. Giang and D. J. Lockhart, unpublished results), although normalization and referencing strategies can be used (D. de Graaf and E. Lander, personal communication).

The second approach avoids PCR altogether and uses multiple rounds of linear amplification based on cDNA synthesis and a template-directed *in vitro* transcription (IVT) reaction^{79–81}. This method has been used to characterize mRNA from single live neurons⁸¹ and even subcellular regions, and more recently to amplify mRNA from 500 to 1,000 cells from microdissected brain tissues for hybridization to spotted cDNA arrays⁸². We have found that the multiple-round cDNA/IVT amplification method produces sufficient quantities of labelled material starting with as little as 1–50 ng total RNA, is highly reproducible (correlation coefficients greater than 0.97), and introduces much less quantitative bias than PCR-based amplification (D. Giang and D. J. Lockhart, unpublished results). These amplification methods facilitate the possibility of monitoring large number of genes starting with very limited amounts of RNA and very few cells. The combination of arrays and powerful amplification strategies promises to be especially important for studies that use human biopsy material from inhomogeneous tissue, and in the areas of developmental biology, immunology and neurobiology.

Genome analysis using arrays

Although nucleic acid arrays are often equated with gene expression analysis, they may be used to collect much of the data that are obtained presently by Southern or northern blot hybridization techniques, but in a more highly parallel fashion (Figs 5, 6). Their utility in polymorphism detection and genotyping is described elsewhere (see review in this issue by Roses, pages 857–865), but there are many additional uses for these versatile tools. For example, genomic DNA samples can be manipulated experimentally to select for particular regions before hybridization to obtain specific types of information. In yeast, the location of hundreds of chromosomal origins of replication can be determined in parallel by enriching for early-replicating

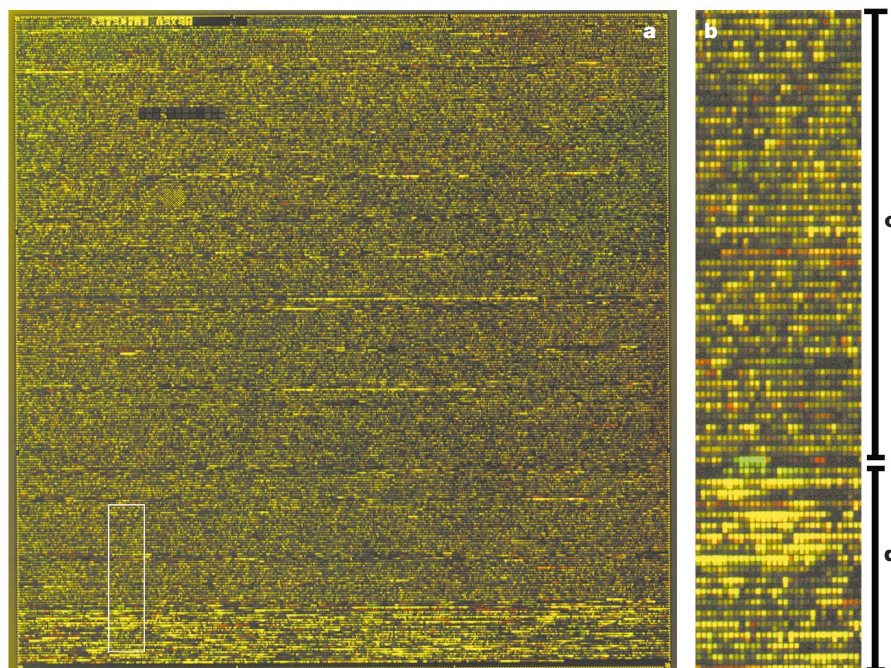
regions using a variation of the Meselson–Stahl procedure and then hybridizing the resulting DNA to full genome arrays (E. A. Winzeler *et al.*, unpublished results). Similarly, as probes for more intergenic regions are synthesized on arrays, it becomes possible to identify protein-binding sites: fragmented chromatin can be crosslinked to a protein and then immunoprecipitated with an antibody to that protein. The DNA fraction of the immunoprecipitate can be labelled and hybridized to identify the approximate location of the binding site. In addition, full genome arrays can be used in the analysis of plasmid libraries in genetic selections such as two-hybrid screens⁸³ or, in principal, for any other type of experiment in which the information is contained in the form of RNA or DNA. Arrays also have applications in biophysical chemistry and biochemistry. For example, single-stranded DNA arrays were converted enzymatically into arrays of double-stranded DNA to characterize the interactions of proteins, and potentially other types of molecules, with double-stranded DNA⁸⁴.

Gene expression and cell circuitry

Is it reasonable to consider the cell as a complex analogue circuit, and to attempt to reverse-engineer the cell circuitry much like an electrical engineer would do by measuring currents and voltages at a variety of nodes and under a variety of input conditions? In the case of the cell, expression levels and expression changes might take the place of electrical measurements, and could be measured under many experimental conditions. Is it possible that a genetic or cellular circuit of reasonable complexity could be adequately decoded or modelled, and if so, how many and what types of measurements and perturbations (or ‘inputs’) would be required so that the problem was not hopelessly underdetermined^{85–89}? Reasonably detailed circuit diagrams can be drawn and simulations of simple genetic circuits have been performed for systems of low complexity (for example, the lytic cycle of phage lambda, and simple control networks in *Escherichia coli* bacteria⁹⁰). But the situation is considerably more complex in the case of a eukaryotic cell. Using yeast as an example, if we assume that the expression level for each gene can be one of only four levels (off, low, medium or high), then if the 6,200 yeast genes behave independently, there are $6,200^4$, or $\sim 1.5 \times 10^{15}$ possible expression states. Of course, the expression levels of different genes are not all independent of one another, and there are some states that are physically unrealistic (for example, all genes ‘off’ or all genes ‘high’), but the number of possible cellular configurations is very large. In addition, coupling between circuit components, the effects of nonlinear feedback, redundancy and even noise and stochastic events make simulating a circuit of this complexity a rather daunting task, and not all relationships and cellular events are reflected at the level of mRNA abundance.

Least clear may be what types of perturbations or inputs are likely to be the most informative in terms of defining the relationships between genes and pathways, and what might be a minimal set of ‘orthogonal perturbations’ (treatments, genetic manipulations or growth conditions that have minimal overlap in their direct cellular effects). Certainly it is possible to delete every yeast gene one at a time (or even several at a time) and measure the expression profile for each mutant strain under a set of different growth conditions^{70,91}. It is also possible to grow yeast on a matrix of thousands of different conditions and measure the resulting expression profiles for a range of mutated strains. It is clear that extensive experiments of this type, combined with information from other measurements such as yeast two-hybrid protein–protein interaction screens⁹², and measurements of protein levels, modification states and cellular localization will lead to useful groupings of genes in terms of function and regulation (that is, a genetic, molecular and functional taxonomy), and to supply some reasonably detailed information about the relationships between certain genes and pathways. In addition, sets of perturbations directed towards specific functions and cellular processes will allow higher-resolution and even mechanistic

Figure 6 Comparative genome hybridization using arrays^{26,106,107}. **a**, Two arrays containing probes to yeast (the complete genome sequence of *S. cerevisiae* strain S288c and some *S. cerevisiae* DNA not present in S288c) were hybridized with fragmented, labelled genomic DNA from two different yeast strains commonly used in genetic studies (W303 and SK1). Red indicates the location of probes that hybridize efficiently only to DNA from the W303 strain, green indicates probes that hybridize only to SK1 DNA, and yellow indicates probes that hybridize equally to the DNA from both strains. **b**, Enlargement of the boxed region in **a**. **c**, Region of the array containing probes to relatively unique protein-coding regions of the genome. **d**, Probes to non-unique regions of the genome (transposable elements, telomeric sequences, transfer RNAs and ribosomal RNAs). Genome regions that are present, absent, or found at higher or lower copy numbers in the two strains are readily detected. The large amount of allelic variation between the strains can be used in mapping studies¹⁰⁸. Related approaches can be used in typing microbial isolates^{29,109} or to identify genetic abnormalities in tumours.



information for significant parts of the overall circuitry^{62,93}. However, given the tremendous complexity of the system, it is unlikely that a complete and detailed cellular circuit diagram will result for even single-celled eukaryotes such as yeast any time in the near future. But that is not to say that construction of even first-order global models and semi-quantitative circuit diagrams is not extremely useful. Such models serve to organize current information, relationships and hypotheses, and can be tremendously helpful for testing new hypotheses, interpreting new observations, designing new experiments and predicting the likely effects of particular chemical, genetic or cellular perturbations. They also serve as a scaffold upon which to build higher-resolution, more quantitative and complete models.

Can we have too much data?

Contrary to what is sometimes thought, the biggest problem for making sense of the extensive results from genomics experiments is not that there is too much data or that there are insufficiently sophisticated algorithms and software tools for querying and visualizing data on this scale. Larger problems of data management and analysis have been solved by airlines, financial institutions, global retailers, high-energy and plasma physicists, the military and global weather predictors, among others. It is often beneficial to have a large number of measurements⁹⁴ and sometimes more data make it possible to analyse results that might otherwise have been too 'messy', and to detect patterns and relationships that would not have been obvious or have sufficient statistical significance with smaller data sets. In many types of studies, it is not possible to control completely all variables, and the individual differences between common sample types may be significant because of experimental difficulties (for example, tissue inhomogeneity or variations in sample procedures) or individual genetic variation (for example, different patients or different tumours). But such factors do not preclude the discovery of some genes that clearly 'cluster' or differentiate between the sample sets. For example, meaningful results can be extracted from the analysis of human tissue collected at different hospitals, by different surgeons and at different times. An essential requirement in these

types of studies is that a sufficient number of experiments be performed across multiple individuals and multiple tissue or tumour samples to account for individual variation and possible tissue inhomogeneity. Furthermore, confidence in the results is increased as conclusions are based on sets of genes that show a consistent response and that are consistently different between two or more sets of results^{49,50,52,53,95}.

Making sense of genomic results

Although the difficulties of sample collection, data collection and experimental design should not be underestimated, one of the most challenging aspects of gene expression analysis is making sense of the vast quantities of data and extracting conclusions and hypotheses that are biologically meaningful. From experiments on global gene expression, we may obtain data for thousands of genes, often forcing us to consider processes, functions and mechanisms about which we know very little. Thus, there is a need for more sophisticated systems of knowledge representation (or 'knowledge bases') that organize the data, facts, observations, relationships and even hypotheses that form the basis of our current scientific understanding. This information needs to be more than just stored; it needs to be available in a way that helps scientists understand and interpret the often complex observations that are becoming increasingly easy to make. Unfortunately, the fact is that the scientific literature has been somewhat haphazardly built, without the benefit of a controlled or restricted vocabulary and a well defined semantic and grammar. To take full advantage of the abilities of the new technologies and the rapidly increasing amount of sequence information it is absolutely essential to incorporate the facts, ideas, connections, observations and so forth, which exist in the scientific literature and in the minds of scientists, into a form that is systematic, organized, linked, visualized and searchable. This clearly requires a great deal of dedicated, systematic human effort, but progress has been made. Databases such as the *Saccharomyces* Genome Database (SGD: genome-www.stanford.edu/Saccharomyces), the Munich Information Center for Protein Sequences (MIPS:

www.mips.biochem.mpg.de), WormBase (www.wormbase.org), the Kyoto Encyclopedia of Genes and Genomes (KEGG: www.genome.ad.jp/kegg), the Encyclopedia of *E. coli* Genes and Metabolism (EcoCyc: http://ecocyc.panbio.com/ecocyc) and FlyBase (flybase.bio.indiana.edu/) incorporate sequence, genetics, gene expression, homology, regulation, function and phenotype information in an organized and useable form^{96–102}. But a step beyond databases of this type are ones in which concepts as well as facts are more fully integrated and related, allowing connections to be made between initially disparate observations and information, and across organisms. It is conceivable that the next step will evolve to the level of a biological 'expert system', not unlike the expert system ('Big Blue') that IBM scientists and engineers built to play chess (successfully) against the world's best chess player. Despite the potential for advancement on this front, it seems unlikely that computational tools will ever replace the trained human brain when it comes to making biological sense of new results. However, the appropriate tools are needed to bring information and relationships to scientist's fingertips so that the most insightful questions can be asked and the most meaningful interpretations made.

Conclusion

For these array-based methods to become truly revolutionary, they must become an integral part of the daily activities of the typical molecular biology laboratory. Despite their impressive and rapidly growing résumé, these technologies are still in their infancy, with plenty of room for technical improvements, further development, and more widespread acceptance and accessibility. We expect that the pattern of development and use of arrays and other parallel genomic methodologies will be similar to that seen for computers and other high-tech electronic devices, which started out as exotic and expensive tools in the hands of the few developers and early adopters, and then moved quickly to become easier to use, more available, less expensive and more powerful, both individually and because of their ubiquity. In fact, nucleic acid array-based methods that previously seemed exotic, and too expensive, are becoming routine as indicated by the huge increase in the number of publications that incorporate data obtained in this way. Despite the relative youth of these approaches, the achievement of technical goals that would have seemed like science fiction only a few years ago is now clearly in view. For example, we expect that measuring the expression level of essentially every gene (including variant splice forms) on an array or two starting with RNA from a small number of cells, or even a single cell, will soon be possible owing to advances in single-cell handling and RNA amplification methods, the output of large-scale sequencing efforts and achievable advances in array technology. In the future, arrays of peptides, proteins, small molecules, mRNAs, clones, tissues, cells and even multicellular organisms such as the nematode worm *Caenorhabditis elegans* may also become common. The combined use of all of these highly parallel methods, along with sequence information, computational tools, integrated knowledge databases, and the traditional approaches of biology, biochemistry, chemistry, physics, mathematics and genetics, increases the hopes of understanding the function and regulation of all genes and proteins, deciphering the underlying workings of the cell, determining the mechanisms of disease, and discovering ways to intervene with or prevent aberrant cellular processes in order to improve human health and well-being. □

- Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
- Hattori, M. *et al.* The DNA sequence of human chromosome 21. *Nature* **405**, 311–319 (2000).
- Lennon, G. G. & Lehrach, H. Hybridization analyses of arrayed cDNA libraries. *Trends Genet.* **7**, 314–317 (1991).
- Kafatos, F. C., Jones, C. W. & Efstratiadis, A. Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. *Nucleic Acids Res.* **7**, 1541–1552 (1979).
- Gillespie, D. & Spiegelman, S. A quantitative assay for DNA-RNA hybrids with DNA immobilized on a membrane. *J. Mol. Biol.* **12**, 829–842 (1965).
- Southern, E. M. *et al.* Arrays of complementary oligonucleotides for analysing the hybridisation behaviour of nucleic acids. *Nucleic Acids Res.* **22**, 1368–1373 (1994).
- Zhao, N., Hashida, H., Takahashi, N., Misumi, Y. & Sakaki, Y. High-density cDNA filter analysis: a

- novel approach for large-scale, quantitative analysis of gene expression. *Gene* **156**, 207–213 (1995).
- Nguyen, C. *et al.* Differential gene expression in the murine thymus assayed by quantitative hybridization of arrayed cDNA clones. *Genomics* **29**, 207–216 (1995).
- Fodor, S. P. A. *et al.* Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767–773 (1991).
- Fodor, S. P. *et al.* Multiplexed biochemical assays with biological chips. *Nature* **364**, 555–556 (1993).
- Pease, A. C. *et al.* Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl Acad. Sci. USA* **91**, 5022–5026 (1994).
- Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
- Shalon, D., Smith, S. J. & Brown, P. O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**, 639–645 (1996).
- DeRisi, J. L., Iyer, V. R. & Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
- Lipshutz, R. J., Fodor, S. P., Gingeras, T. R. & Lockhart, D. J. High density synthetic oligonucleotide arrays. *Nature Genet.* **21**, 20–24 (1999).
- Bowtell, D. D. Options available—from start to finish—for obtaining expression data by microarray. *Nature Genet.* **21**, 25–32 (1999).
- Edman, C. F. *et al.* Electric field directed nucleic acid hybridization on microchips. *Nucleic Acids Res.* **25**, 4907–4914 (1997).
- Sosnowski, R. G., Tu, E., Butler, W. F., O'Connell, J. P. & Heller, M. J. Rapid determination of single base mismatch mutations in DNA hybrids by direct electric field control. *Proc. Natl Acad. Sci. USA* **94**, 1119–1123 (1997).
- Gray, D. E., Case-Green, S. C., Fell, T. S., Dobson, P. J. & Southern, E. M. Ellipsometric and interferometric characterization of DNA probes immobilised on a combinatorial array. *Langmuir* **13**, 2833–2842 (1997).
- Walt, D. R. Bead-based fiber-optic arrays. *Science* **287**, 451 (2000).
- Michael, K. L., Taylor, L. C., Schultz, S. L. & Walt, D. R. Randomly ordered addressable high-density optical sensor arrays. *Anal. Chem.* **70**, 1242–1248 (1998).
- Ferguson, J. A., Boles, T. C., Adams, C. P. & Walt, D. R. A fiber-optic DNA biosensor microarray for the analysis of gene expression. *Nature Biotechnol.* **14**, 1681–1684 (1996).
- Spellman, P. T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
- Cho, R. J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65–73 (1998).
- The Chipping forecast. *Nature Genet.* **21** (Suppl.), 1–60 (1999).
- Wodicka, L., Dong, H., Mittmann, M., Ho, M.-H. & Lockhart, D. J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol.* **15**, 1359–1367 (1997).
- White, K. I., Rifkin, S. A., Hurlbut, P. & Hogness, D. S. Microarray analysis of *Drosophila* development during metamorphosis. *Science* **286**, 2179–2184 (1999).
- Chambers, J. *et al.* DNA microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression. *J. Virol.* **73**, 5757–5766 (1999).
- Gingeras, T. R. *et al.* Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic mycobacterium DNA arrays. *Genome Res.* **8**, 435–448 (1998).
- Lockhart, D. J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675–1680 (1996).
- Liang, P. & Pardee, A. B. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**, 967–971 (1992).
- Shinkets, R. A. *et al.* Gene expression analysis by transcript profiling coupled to a gene database query. *Nature Biotechnol.* **17**, 798–803 (1999).
- Ivanova, N. B. & Belyavsky, A. V. Identification of differentially expressed genes by restriction endonuclease-based gene expression fingerprinting. *Nucleic Acids Res.* **23**, 2954–2958 (1995).
- Kato, K. Description of the entire mRNA population by a 3' end cDNA fragment generated by class IIS restriction enzymes. *Nucleic Acids Res.* **23**, 3685–3690 (1995).
- Bachem, C. W. *et al.* Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. *Plant J.* **9**, 745–753 (1996).
- Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
- Boucherie, H. *et al.* Two-dimensional protein map of *Saccharomyces cerevisiae*: construction of a gene-protein index. *Yeast* **11**, 601–613 (1995).
- Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.* **17**, 994–999 (1999).
- Mann, M. Quantitative proteomics? *Nature Biotechnol.* **17**, 954–955 (1999).
- Oda, Y., Huang, K., Cross, F. R., Cowburn, D. & Chait, B. T. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl Acad. Sci. USA* **96**, 6591–6596 (1999).
- Burns, N. *et al.* Large-scale analysis of gene expression, protein localization, and gene disruption in *Saccharomyces cerevisiae*. *Genes Dev.* **8**, 1087–1105 (1994).
- Ross-Macdonald, P., Sheehan, A., Roeder, G. S. & Snyder, M. A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **94**, 190–195 (1997).
- Ross-Macdonald, P. *et al.* Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402**, 413–418 (1999).
- Niedenthal, R. K., Riles, L., Johnston, M. & Hegemann, J. H. Green fluorescent protein as a marker for gene expression and subcellular localization in budding yeast. *Yeast* **12**, 773–786 (1996).
- Zong, Q., Schummer, M., Hood, L. & Morris, D. R. Messenger RNA translation state: the second dimension of high-throughput expression screening. *Proc. Natl Acad. Sci. USA* **96**, 10632–10636 (1999).
- Johannes, G., Carter, M. S., Eisen, M. B., Brown, P. O. & Sarnow, P. Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. *Proc. Natl Acad. Sci. USA* **96**, 13118–13123 (1999).
- Diehn, M., Eisen, M. B., Botstein, D. & Brown, P. O. Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. *Nature Genet.* **25**, 58–62 (2000).
- Weinstein, J. N. Fishing expeditions. *Science* **282**, 628–629 (1998).
- Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
- Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression

- profiling. *Nature* **403**, 503–510 (2000).
51. Mack, D. H. *et al.* in *Deciphering Molecular Circuitry Using High-Density DNA Arrays* (eds Hihich, E. & Croce, E.) 85–108 (Plenum, New York, 1998).
52. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96**, 6745–6750 (1999).
53. Perou, C. M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* **96**, 9212–9217 (1999).
54. Ross, D. T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.* **24**, 227–235 (2000).
55. Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* **24**, 236–244 (2000).
56. Fambrough, D., McClure, K., Kazlauskas, A. & Lander, E. S. Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes. *Cell* **97**, 727–741 (1999).
57. Lee, S. B. *et al.* The Wilms tumor suppressor WT1 encodes a transcriptional activator of amphiregulin. *Cell* **98**, 663–673 (1999).
58. Harkin, D. P. *et al.* Induction of GADD45 and JNK/SAPK-dependent apoptosis following inducible expression of BRCA1. *Cell* **97**, 575–586 (1999).
59. Mei, R. *et al.* Genome-wide detection of allelic imbalance using human SNPs and high density DNA arrays. *Genome Res.* (in the press).
60. Pollack, J. R. *et al.* Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genet.* **23**, 41–46 (1999).
61. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet.* **20**, 207–211 (1998).
62. Holstege, F. C. *et al.* Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95**, 717–728 (1998).
63. Wyrick, J. J. *et al.* Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. *Nature* **402**, 418–421 (1999).
64. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).
65. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
66. Wen, X. *et al.* Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA* **95**, 334–339 (1998).
67. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998).
68. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285 (1999).
69. Wolfsberg, T. G. *et al.* Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*. *Genome Res.* **9**, 775–792 (1999).
70. Marton, M. J. *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* **4**, 1293–1301 (1998).
71. Gray, N. S. *et al.* Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* **281**, 533–538 (1998).
72. Rosania, G. R. *et al.* Myosoverin: a microtubule binding molecule with novel cellular effects. *Nature Biotechnol.* **18**, 304–308 (2000).
73. Emmert-Buck, M. R. *et al.* Laser capture microdissection. *Science* **274**, 998–1001 (1996).
74. Bonner, R. F. *et al.* Laser capture microdissection: molecular analysis of tissue. *Science* **278**, 1481–1483 (1997).
75. Simone, N. L., Bonner, R. F., Gillespie, J. W., Emmert-Buck, M. R. & Liotta, L. A. Laser-capture microdissection: opening the microscopic frontier to molecular analysis. *Trends Genet.* **14**, 272–276 (1998).
76. Wang, A. M., Doyle, M. V. & Mark, D. F. Quantitation of mRNA by the polymerase chain reaction. *Proc. Natl Acad. Sci. USA* **86**, 9717–9721 (1989).
77. Dulac, C. Cloning of genes from single neurons. *Curr. Top. Dev. Biol.* **36**, 245–258 (1998).
78. Jena, P. K., Liu, A. H., Smith, D. S. & Wysocki, L. J. Amplification of genes, single transcripts and cDNA libraries from one cell and direct sequence analysis of amplified products derived from one molecule. *J. Immunol. Methods* **190**, 199–213 (1996).
79. Kwok, D. Y. *et al.* Transcription-based amplification system and detection of amplified human immunodeficiency virus type 1 with a bead-based sandwich hybridization format. *Proc. Natl Acad. Sci. USA* **86**, 1173–1177 (1989).
80. Guatelli, J. C. *et al.* Isothermal, in vitro amplification of nucleic acids by a multienzyme reaction modeled after retroviral replication. *Proc. Natl Acad. Sci. USA* **87**, 7797 (1990).
81. Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc. Natl Acad. Sci. USA* **89**, 3010–3014 (1992).
82. Luo, L. *et al.* Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nature Med.* **5**, 117–122 (1999).
83. Cho, R. J. *et al.* Parallel analysis of genetic selections using whole genome oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **95**, 3752–3757 (1998).
84. Bulyk, M. L., Gentale, E., Lockhart, D. J. & Church, G. M. Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nature Biotechnol.* **17**, 573–577 (1999).
85. Brent, R. Genomic biology. *Cell* **100**, 169–183 (2000).
86. McAdams, H. H. & Shapiro, L. Circuit simulation of genetic networks. *Science* **269**, 650–656 (1995).
87. McAdams, H. H. & Arkin, A. It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet.* **15**, 65–69 (1999).
88. Bhalla, U. S. & Iyengar, R. Emergent properties of networks of biological signaling pathways. *Science* **283**, 381–387 (1999).
89. Weng, G., Bhalla, U. S. & Iyengar, R. Complexity in biological signaling systems. *Science* **284**, 92–96 (1999).
90. Arkin, A., Ross, J. & McAdams, H. H. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* **149**, 1633–1648 (1998).
91. Winzler, E. *et al.* Functional characterization of the *Saccharomyces cerevisiae* genome by precise deletion and parallel analysis. *Science* **285**, 901–906 (1999).
92. Fields, S. & Song, O. A novel genetic system to detect protein–protein interactions. *Nature* **340**, 245–246 (1999).
93. Roberts, C. J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
94. Brown, P. O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21**, 33–37 (1999).
95. Ly, D., Lockhart, D. J., Lerner, R. & Schultz, P. G. Mitotic misregulation and human aging. *Science* **287**, 2486–2492 (2000).
96. Cherry, J. M. *et al.* Genetic and physical maps of *Saccharomyces cerevisiae*. *Nature* **387**, 67–73 (1997).
97. Ball, C. A. *et al.* Integrating functional genomic information into the *Saccharomyces* Genome Database. *Nucleic Acids Res.* **28**, 77–80 (2000).
98. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **28**, 37–40 (2000).
99. Walsh, S., Anderson, M. & Cartinhour, S. W. ACEDB: a database for genome information. *Methods Biochem. Anal.* **39**, 299–318 (1998).
100. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
101. The FlyBase Consortium. The FlyBase Database of the *Drosophila* Genome Projects and community literature. *Nucleic Acids Res.* **27**, 85–88 (1999).
102. Karp, P. D., Riley, M., Paley, S. M., Pellegrini-Toole, A. & Krummenacker, M. EcoCyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.* **27**, 50–53 (1999).
103. Iyer, V. & Struhl, K. Absolute mRNA levels and transcriptional initiation rates in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **93**, 5208–5212 (1996).
104. Lee, C. K., Klopp, R. G., Weindrich, R. & Prolla, T. A. Gene expression profile of aging and its retardation by caloric restriction. *Science* **285**, 1390–1393 (1999).
105. Fan, J.-B. *et al.* Parallel genotyping of human SNPs using generic oligonucleotide tag arrays. *Genome Res.* (in the press).
106. Lashkari, D. A. *et al.* Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl Acad. Sci. USA* **94**, 13057–13062 (1997).
107. Winzler, E., Lee, B., McCusker, J. & Davis, R. Whole genome genetic typing using high-density oligonucleotide arrays. *Parasitology* **118**, S73–S80 (1999).
108. Winzler, E. A. *et al.* Direct allelic variation scanning of the yeast genome. *Science* **281**, 1194–1197 (1998).
109. Troesch, A. *et al.* Mycobacterium species identification and rifampin resistance testing with high-density DNA probe arrays. *J. Clin. Microbiol.* **37**, 49–55 (1999).

Acknowledgements

We thank S. Fodor, M. Chee, R. Davis, L. Stryer, E. Lander, H. Dong, L. Wodicka, R. Cho, D. Giang, P. Zarrinkar, C. Barlow, J. Gentry, P. Schultz and R. Abagyan for their on-going help and patience, and B. Geierstanger, G. Hampton and S. Kay for helpful comments and a critical reading of the manuscript.

Proteomics to study genes and genomes

Akhilesh Pandey* & Matthias Mann†

*Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, and Department of Pathology, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA

†Protein Interaction Laboratory (PIL), University of Southern Denmark, Campusvej 55, DK-5230 Odense M, and Denmark and Protana A/S, Staermosegaardsvej 16, DK-5230 Odense M, Denmark (e-mail: mann@cebi.sdu.dk)

Proteomics, the large-scale analysis of proteins, will contribute greatly to our understanding of gene function in the post-genomic era. Proteomics can be divided into three main areas: (1) protein micro-characterization for large-scale identification of proteins and their post-translational modifications; (2) 'differential display' proteomics for comparison of protein levels with potential application in a wide range of diseases; and (3) studies of protein-protein interactions using techniques such as mass spectrometry or the yeast two-hybrid system. Because it is often difficult to predict the function of a protein based on homology to other proteins or even their three-dimensional structure, determination of components of a protein complex or of a cellular structure is central in functional analysis. This aspect of proteomic studies is perhaps the area of greatest promise. After the revolution in molecular biology exemplified by the ease of cloning by DNA methods, proteomics will add to our understanding of the biochemistry of proteins, processes and pathways for years to come.

Large-scale DNA sequencing has transformed biomedical research in a short span of time. With the discovery of most human genes, it is now apparent that a 'factory approach' to address biological problems is desirable if we are to gain a comprehensive understanding of complex biological processes. In this article we will review how proteomics is similarly making a crucial contribution to our understanding of biology and medicine through the global analysis of gene products.

Defining proteomics

Proteomics is the large-scale study of proteins, usually by biochemical methods. The word proteomics has been associated traditionally with displaying a large number of proteins from a given cell line or organism on two-dimensional polyacrylamide gels¹⁻⁴. In this sense proteomics already dates back to the late 1970s when researchers started to build databases of proteins using the then newly developed technique of two-dimensional gel electrophoresis⁵ (Box 1). This resulted in extensive cataloguing of spots from two-dimensional gels to create databases of all expressed proteins. However, even when such gels could be run reproducibly between laboratories, determining the identity of the proteins was difficult because of a lack of sensitive and rapid analytical methods for protein characterization (such as the polymerase chain reaction and the automated sequencer for DNA analysis). In the 1990s, biological mass spectrometry emerged as a powerful analytical method that removed most of the limitations of protein analysis. This development, coupled with the availability of the entire human coding sequence in public databases, marks the beginning of a new era. Today, the term proteomics covers much of the functional analysis of gene products or 'functional genomics', including large-scale identification or localization studies of proteins and interaction studies using the yeast two-hybrid system. The more focused large-scale study of protein structure, however, is

usually not included and designated 'structural genomics' instead⁶. Likewise, strategies that target only genes or messenger RNA, such as large-scale mutagenesis or antisense experiments, should not be considered part of proteomics.

Why is proteomics necessary?

With the accumulation of vast amounts of DNA sequences in databases, researchers are realizing that merely having complete sequences of genomes is not sufficient to elucidate biological function. A cell is normally dependent upon a multitude of metabolic and regulatory pathways for its survival. There is no strict linear relationship between genes and the protein complement or 'proteome' of a cell. Proteomics is complementary to genomics because it focuses on the gene products, which are the active agents in cells. For this reason, proteomics directly contributes to drug development as almost all drugs are directed against proteins.

The existence of an open reading frame (ORF) in genomic data does not necessarily imply the existence of a functional gene. Despite the advances in bioinformatics, it is still difficult to predict genes accurately from genomic data (see review in this issue by Eisenberg *et al.*, pages 823–826, and refs 7, 8). Although the sequencing of related organisms will ease the problem of gene prediction through comparative genomics, the success rate for correct prediction of the primary structure is still low^{9,10}. This is particularly true in the case of small genes (which can be missed entirely) or genes with little or no homology to other known genes. A recent study concluded that the error rate was at least 8% in the annotations for 340 genes from the *Mycoplasma genitalium* genome¹¹. If such error rates are extrapolated to the human genome, the outcome and consequences can easily be imagined. Therefore, verification of a gene product by proteomic methods is an important first step in 'annotating the genome'. Modifications of the proteins that are not apparent from the DNA sequence, such as isoforms and

post-translational modifications, can be determined only by proteomic methodologies. Furthermore, it may be necessary to determine the protein expression level directly as mRNA levels may or may not correlate with the protein level^{12,13}. The localization of gene products, which is often difficult to predict from the sequence, can be determined experimentally. Mechanisms such as regulation of protein function by proteolysis, recycling and sequestration in cell compartments affect gene products and not genes. Finally, protein–protein interactions and the molecular composition of cellular structures such as organelles can be determined only at the protein level.

Identification and analysis of proteins

Protein preparation methods

One of the most crucial steps in proteomics is obtaining and handling the protein sample. Out of the entire complement of the genome of about 100,000 genes, a given cell line may express about 10,000 genes and an even higher number is expressed in tissues. Furthermore, the dynamic range of abundance of proteins in biological samples can be as high as 10^6 . Because even the best two-dimensional gels can routinely resolve no more than 1,000 proteins, it is obvious that only the most abundant proteins can be visualized by gel electrophoresis if a crude protein mixture is used. The ideal solution to reduce complexity and differences in abundance is to use affinity-based protein purification strategies using the whole protein complement. For example, the erythropoietin receptor is of medium abundance, occurring in about 1,000 copies per cell, or less than two picomoles (100 ng) in one litre of cell culture. This protein would not be visualized from whole-cell extracts but can be enriched easily by antibody-based affinity purification to yield a silver-stained band. This fact has to be borne in mind if signalling and other regulatory molecules are being studied.

After obtaining the protein fraction, the method of choice for proteomic studies is one- or two-dimensional gel electrophoresis. The advantages of one-dimensional electrophoresis as a preparation method are that virtually all proteins are soluble in SDS, the range of relative molecular mass from 10,000 to 300,000 is readily covered, and extremely acidic and basic proteins are easily visualized.

Mass spectrometric identification of proteins

The most significant breakthrough in proteomics has been the mass spectrometric identification of gel-separated proteins, which extends analysis far beyond the mere display of proteins. Mass spectrometry has essentially replaced the classical technique of Edman degradation even in traditional protein chemistry, because it is much more sensitive, can deal with protein mixtures and offers much higher throughput. It relies on digestion of gel-separated proteins into peptides by a sequence-specific protease such as trypsin. The reason for analysing peptides rather than proteins is that gel-separated proteins are difficult to elute and to analyse by mass spectrometry, and that the molecular weight of proteins is not usually sufficient for database identification. In contrast, peptides are easily eluted from gels and even a small set of peptides from a protein provides sufficient information for identification. The steps typically involved in the mass spectrometric analysis of a protein are illustrated by an example that shows analysis of a molecule involved in platelet-derived growth factor (PDGF) signalling (Fig. 1). A detailed protocol describing methods and strategies for the mass spectrometric identification of signalling molecules can be found in ref. 14.

There are two main approaches to mass spectrometric protein identification. In the 'peptide-mass mapping' approach, initially suggested by Henzel and co-workers¹⁵, the mass spectrum of the eluted peptide mixture is acquired, which results in a 'peptide-mass fingerprint' of the protein being studied. This mass spectrum is obtained by a relatively simple mass spectrometric method — matrix-assisted laser desorption/ionization (MALDI) — which results in a time-of-flight distribution of the peptides comprising the mixture (Box 2 and Fig. 1b). Advances have been made in automation

Box 1

Defining proteomics

Proteomics – the classical definition

- Two-dimensional gels of cell lysates and annotation
- Two-dimensional gels to visualize differential protein expression

Proteomics – in the post-genomics era

- Protein identification:
 - One-dimensional gels (for example, analysis after affinity purification)
 - Two-dimensional gels (for example, analysis after affinity purification, body fluids, etc.)
 - Protein chips (chips coated with, for example, proteins or antibodies)
 - Proteins/protein complexes in solution (identification without electrophoresis)
- Post-translational modifications
 - Phosphorylation
 - Glycosylation
- Determining Function
 - Assays for enzymatic activity or determining substrates⁷⁵
 - Bioassays for cytokines, receptor/ligand-binding assays
 - Localization within the cells (GFP fusions)
 - Proteomic analysis using large-scale mouse knockouts⁷⁶ or RNA interference⁷⁷.
 - Phenotypic analysis using deletion strains⁷⁸
- Molecular Medicine (no longer just pharmaceuticals)
 - Finding molecular (protein) drug targets
 - Disrupting protein–protein interactions using drugs
 - Large-scale animal assays for recombinant proteins, antibodies and inhibitors
- Differential display by two-dimensional gels (superseded by DNA-based array in many situations)
 - Limited applications in:
 - Body fluids (for example, serum and urine)
 - Variants resulting from post-translational modifications
- Protein–protein interaction
 - Direct DNA readout
 - Yeast two hybrid
 - Phage display
 - Ribosome display⁷⁹
 - RNA–peptide fusions⁸⁰
- Protein identification
 - Affinity purification and mass spectrometry

of the MALDI identification procedure whereby hundreds of protein spots can be excised, digested enzymatically, their mass spectra obtained and automatically searched against databases^{16,17}. As more full-length human genes are represented in the database, the success rate of identification by MALDI will increase further.

In a two-step procedure for rapid and unambiguous protein identification, MALDI fingerprinting is the first step¹⁸. The second method for protein identification relies on fragmentation of individual peptides in the mixture to gain sequence information. In this method, the peptides are ionized by 'electrospray ionization' directly from the liquid phase. The peptide ions are sprayed into a 'tandem mass spectrometer' which has the ability to resolve peptides in a mixture, isolate one species at a time and dissociate it into amino- or carboxy-terminal-containing fragments (Fig. 1c). The tandem mass spectrometric method is technically more complex and less scalable than MALDI fingerprinting. Its main advantage is that sequence information derived from several peptides is much more specific for the identification of a protein than a list of peptide masses. The fragmentation data can not only be used to search protein sequence databases but also nucleotide databases such as expressed sequence

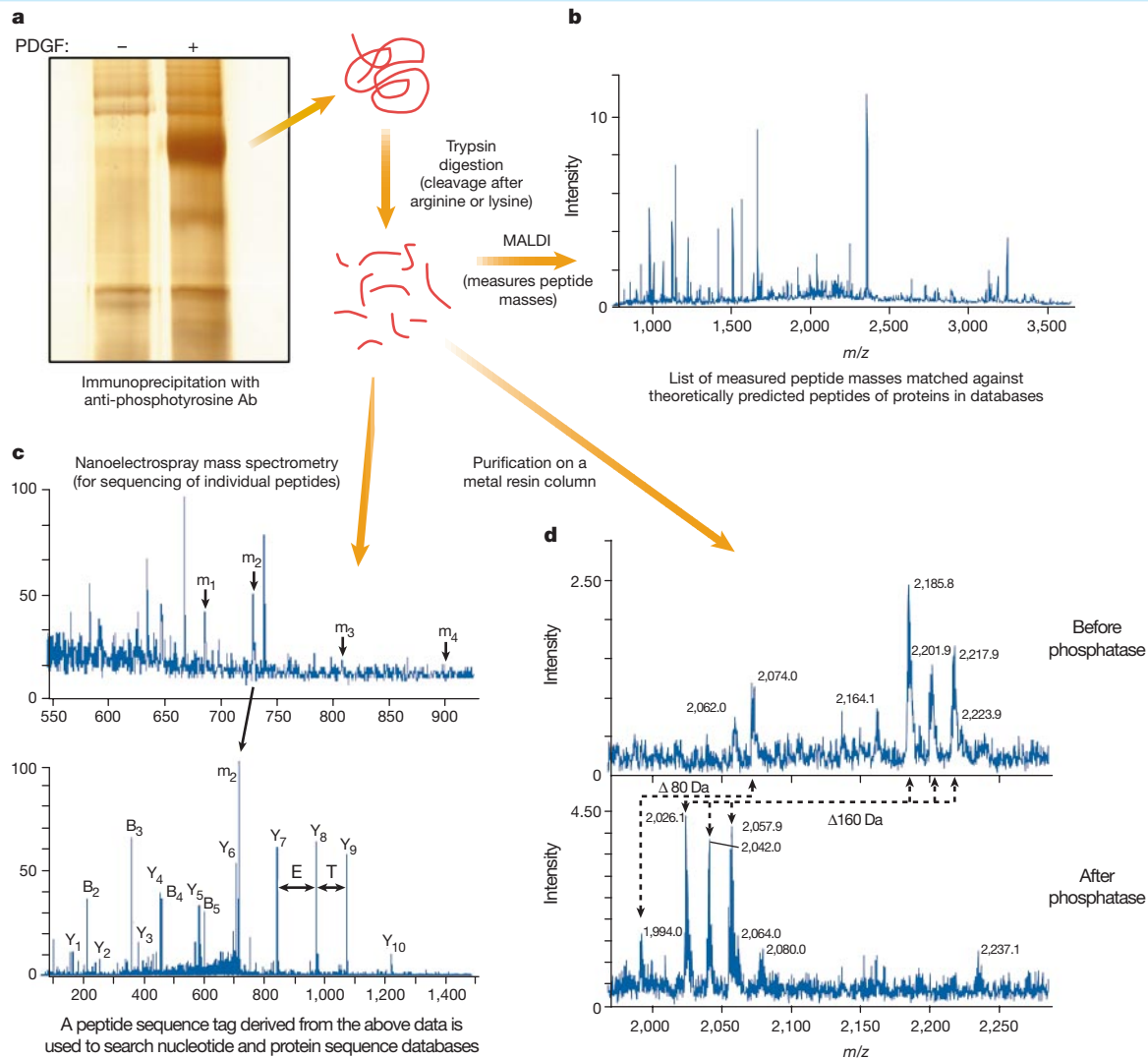


Figure 1 A strategy for mass spectrometric identification of proteins and post-translational modifications. **a**, Responsive cells such as NIH 3T3 fibroblasts are treated with PDGF followed by immunoprecipitation of cell lysates with anti-phosphotyrosine antibodies. After one-dimensional gel electrophoresis, the gel is silver stained, the protein band excised as shown and subjected to digestion with trypsin. This results in peptides with arginine or lysine at their C termini as a result of the cleavage specificity of trypsin. **b**, An aliquot of the supernatant containing tryptic peptides is analysed by MALDI, which results in a peptide-mass fingerprint of the protein. **c**, The remainder of the supernatant is desalted and analysed by nanoelectrospray tandem mass spectrometry. The top panel shows the individual peptide peaks in the mass spectrum. The bottom panel shows how sequence can be derived by fragmentation of the chosen peptide (m_2) by tandem mass spectrometry. **d**, The phosphopeptides may be enriched by purifying the peptide mixture over a metal resin microcolumn. The resulting peptides can then be analysed by MALDI as shown (and subsequently by nanoelectrospray) before and after treatment with alkaline phosphatase. The panel shows a singly phosphorylated (showing a shift of 80 Da) and a doubly phosphorylated (showing a shift of 160 Da) peptide in the MALDI spectrum. (Fig. 1d courtesy of O. N. Jensen and A. Stensballe.)

tag (EST) databases and more recently even raw genomic sequence databases (B. Küster, P. Mortensen, J. S. Andersen and M. Mann, unpublished data).

New developments in mass spectrometry

Biological mass spectrometry is still evolving rapidly owing to continued technological advances in various areas. For instance, a new type of mass spectrometer that combines a MALDI ion source with a highly efficient tandem mass spectrometer unit that can fragment the individual peptides has recently been developed¹⁹. If this 'MALDI quadrupole time of flight' instrument proves to be sufficiently sensitive, it would combine the high throughput of the peptide mapping method with the specificity of the peptide sequencing method, allowing a one-step instead of a two-step mass spectrometric analysis strategy. In our experience, this instrument already significantly improves the analysis of small proteins and improves the

throughput when analysing simple protein mixtures. There are also efforts at miniaturizing protein preparation using microfabricated 'chips', which have obtained promising results^{20–22}. However, these methods have not yet yielded the sensitivity or robustness of preparations using standard tube or microtitre plate formats. There are also longstanding efforts to scan one- or two-dimensional gels directly by MALDI mass spectrometry^{23,24}. A recent variation uses an intercalating membrane containing immobilized trypsin for digestion of proteins during electrophoretic transfer onto a collecting membrane. The membrane is then rasterized and analysed by MALDI yielding a peptide map for each position of the gel^{25,26}.

In the future, it would be desirable to analyse a protein sample directly by mass spectrometry, without gel separation or enzymatic digestion. Smith *et al.* have loaded crude protein extract into a capillary and performed capillary electrophoresis to separate the proteins

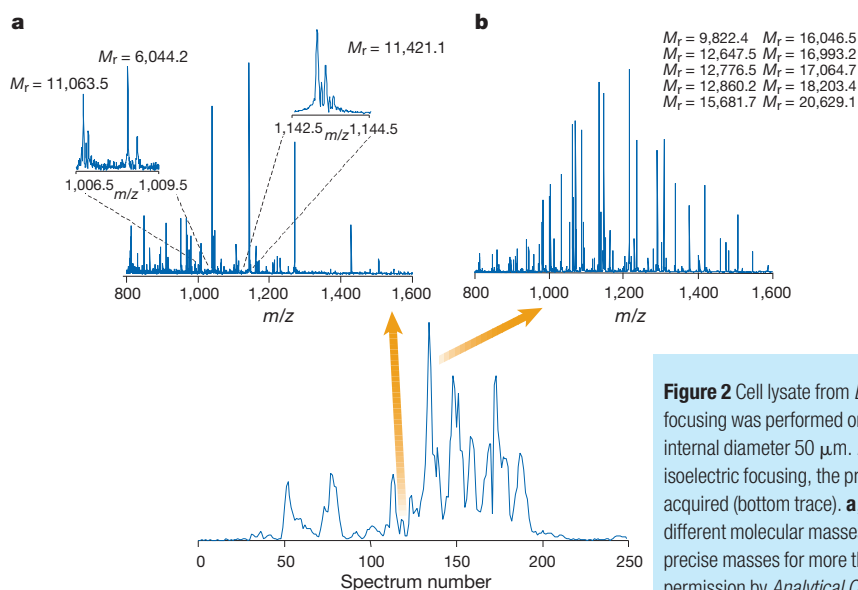


Figure 2 Cell lysate from *Escherichia coli* analysed by FTICR. Capillary isoelectric focusing was performed on ~300 ng *E. coli* total cell lysate in a coated capillary of internal diameter 50 μm . *E. coli* were grown in isotopically depleted medium. After isoelectric focusing, the proteins were eluted into the mass spectrometer and spectra acquired (bottom trace). **a**, High-resolution spectrum for charge states representing different molecular masses present in a single scan. **b**, Mass spectrum showing precise masses for more than ten co-eluting protein species. (Reprinted with permission by *Analytical Chemistry*.)

by their isoelectric point²⁷. The separated proteins were then infused directly into a specialized Fourier-transformed ion cyclotron resonance (FTICR) mass spectrometer (Fig. 2), and the precise molecular masses of hundreds of proteins were acquired during a single run. In this experiment, the mass distribution was biased towards small proteins and only the masses, not the identity of the proteins, were determined. But in the future it may become possible to use this strategy to identify proteins by on-line fragmentation of the proteins^{28,29}. This would enable researchers to perform the whole proteomic analysis in a single automated experiment at least for a subset of soluble proteins of medium abundance.

Post-translational modifications

One of the unique features of proteomics studies is the ability to analyse the post-translational modifications of proteins. Phosphory-

lation, glycosylation and sulphation as well as many other modifications are extremely important for protein function as they can determine activity, stability, localization and turnover. These modifications are not generally apparent from genomic sequence or mRNA expression data. Whereas mass spectrometry is the proteomic method of choice to determine protein modifications, this task is much more difficult than the mere determination of protein identity. Minimal data is sufficient to identify the protein in sequence databases — often as few as one or two peptides need to be fragmented. However, for obtaining the nature and location of post-translational modifications, all the peptides that do not have the expected molecular mass need to be analysed further. Because of this and other reasons, much more material is needed to study post-translational modifications than is required for protein

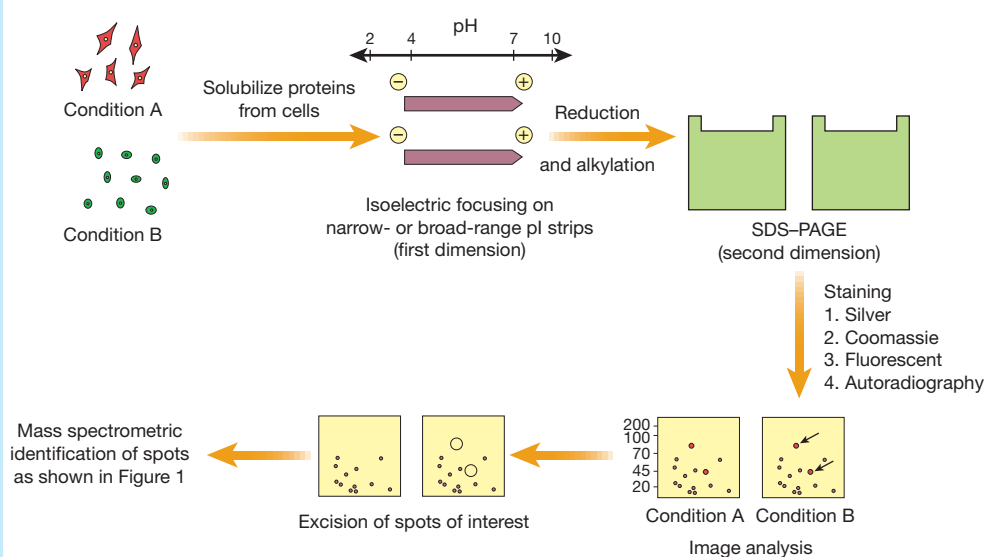


Figure 3 A schematic showing the two-dimensional gel approach. Cells (or tissue) derived from two different conditions, A and B, are harvested and the proteins solubilized. The crude protein mixture is then applied to a 'first dimension' gel strip that separates the proteins based on their isoelectric points. After this step, the strip is subjected to reduction and alkylation and applied to a 'second dimension' SDS-PAGE gel where proteins are denatured and separated on the basis of size. The gels are then fixed and the proteins visualized by silver staining. Silver staining is less quantitative than Coomassie blue but more sensitive and is also compatible with mass spectrometric analysis. After staining, the resulting protein spots are recorded and quantified. Image analysis requires sophisticated software and remains one of the most labour-intensive parts of the two-dimensional gel approach. The spots of interest are then excised and subjected to mass spectrometric analysis.

identification. Continuing progress is being made in this field, especially in the case of phosphorylation. Phosphorylation events can be studied by generic strategies, because phosphopeptides are 80 Da heavier than their unmodified counterparts, give rise to a specific fragment (PO_3^- , mass 79), bind to metal resins, are recognized by specific antibodies and the phosphate groups can be removed by phosphatases^{30–34}. As an example, Fig 1d shows the detection of phosphopeptides following metal resin-based affinity micropurification and phosphatase treatment.

Phosphorylation and signalling pathways

Several receptor-mediated signalling pathways result in tyrosine phosphorylation of a large set of substrates. To identify these substrates, the lysates from unstimulated and growth factor-stimulated cells can be prepared and resolved by two-dimensional gels. The proteins of interest can be detected by ^{32}P labelling or by western blotting with antibodies that recognize only the activated state of molecules (such as phosphotyrosine- or phosphoserine-specific antibodies). These spots can then be identified by mass spectrometry as demonstrated recently³⁵. A better alternative, however, is to first enrich for these substrates by using anti-phosphotyrosine antibodies in an immunoprecipitation step followed by mass spectrometric identification. Several known and new components were recently reported in one such study on the epidermal growth factor (EGF)-receptor pathway³⁶.

Differential-display proteomics

The two-dimensional gel approach

Until recently, proteomics was almost synonymous with two-dimensional gel electrophoresis (Fig. 3). In biomedical applications of the comparative two-dimensional gel approach, the objective is usually to identify proteins that are up- or downregulated in a disease-specific manner for use as diagnostic markers or therapeutic targets. There are several technical challenges in such experiments. First, hydrophobic and large proteins usually do not enter the second

dimension of the gel. Second, the issue of dynamic range makes it difficult to visualize all but the most abundant proteins. Particularly in body fluids such as serum and cerebrospinal fluid, more than 99% of the protein complement consists of serum albumin and globulins. Third, because of the biological variation inherent in these samples, it is difficult to define normal protein-expression patterns that can be compared with the disease state. For several of these applications, methods of array-based mRNA expression profiling can not only be more comprehensive (as they provide data on all the genes applied to the chip), but also faster and more convenient, as shown by a number of studies (see review in this issue by Lockhart and Winzler, pages 827–836, and refs 37–40).

In spite of these difficulties of comparing two-dimensional gel patterns, several applications have appeared in the literature. For example, Celis and co-workers have found a putative urinary marker, psoriasin, which can be used for the follow-up of patients with bladder squamous cell carcinomas⁴¹. This marker was identified when they compared the profile of secreted proteins from normal tissue with that from cancerous tissue. A similar study compared the proteome of normal human luminal and myoepithelial breast cells using immunopurified cell populations. It detected 170 protein spots that were twofold differentially expressed⁴², of which 51 were identified. However, almost all of these proteins were abundant cytoskeletal proteins such as actin and keratin. A recent study compared the protein complement from different fractions of brain extracts from two different strains of mice⁴³, finding over 1,000 genetically variant protein spots. Such studies may be useful in other situations as well, for example, in comparing the proteome of wild-type with that of knockout mice. Toxicology studies frequently use proteomic analysis to understand the mechanism of action of a drug or to identify its targets. Aicher and colleagues discovered an association between decreased levels of a calcium-binding protein, calbindin-D 28K, and cyclosporine A-induced nephrotoxicity when kidney samples were compared from species that were either

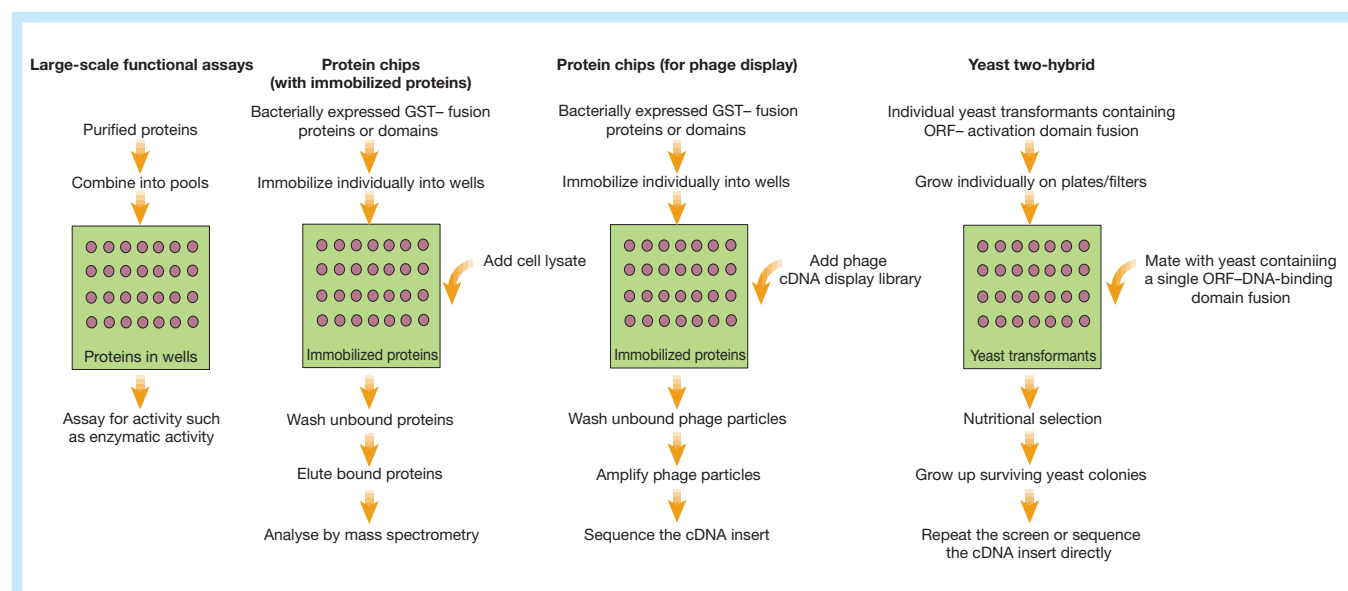


Figure 4 A schematic showing use of arrays for proteomic analysis. Recombinant proteins can be expressed and purified in a large-scale format. These proteins are pooled into wells as shown and assayed for functions such as enzymatic activity. This approach has been termed biochemical genomics. A protein chip can be prepared in several ways. The surface can be immobilized with recombinant proteins or their domains (such as bacterially expressed GST-fusion proteins) and then cell lysates containing putative interaction partners are applied to the chip followed by washing to remove unbound material. The bound proteins can then be eluted and identified by mass spectrometry. Alternatively, instead of cell lysates, a phage cDNA display library can be applied to the chip followed by washing and amplification steps to isolate individual interacting phage particles. The inserts in these phage particles can then be sequenced to determine the identity of the interacting partners. The yeast two-hybrid system is also amenable to an array-based analysis. First, yeast cells can be transformed with individual ORF-activation domain fusions. These cells can be grown in an array format on plates or filters such that each element of the array contains a yeast clone with a unique ORF. Such an array can be probed in a mating assay with yeast cells containing a single ORF-DNA-binding domain fusion, one at a time. The nutritional selection ensures that only the yeast cells containing interacting partners survive. These interacting clones can be re-screened to reduce false positives or be sequenced directly.

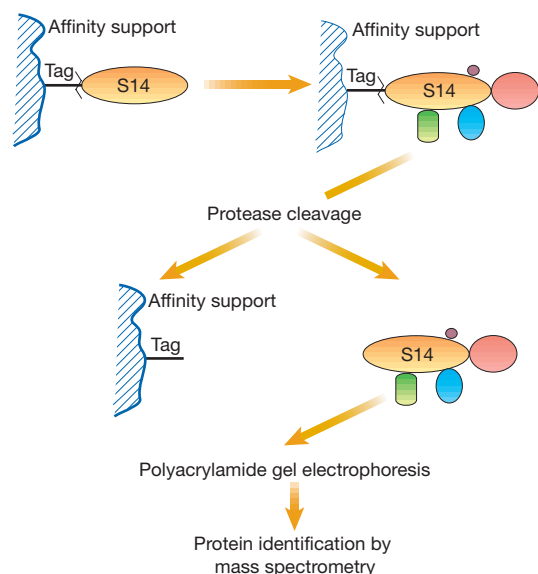


Figure 5 A generic strategy to isolate interacting proteins. The protein of interest is expressed as a fusion protein with a cleavable affinity tag to identify interacting proteins. In this case, S14 protein (spot S14 identified from gel shown in Fig. 6a) is immobilized onto agarose beads using a GST tag. Nuclear cell extracts are incubated with the beads and the beads washed extensively. Thrombin is used to cleave between the GST and the S14 protein, which results in elution of all proteins that are specifically bound to S14. The advantage of this method is that the proteins that are nonspecifically bound to the matrix or the tag itself are not eluted. The eluted proteins are resolved by one- or two-dimensional gel electrophoresis and compared to GST alone. The bands or spots corresponding to proteins specifically bound to the tagged proteins are excised and analysed by mass spectrometry. (Figure courtesy of A. King)

susceptible or resistant to nephrotoxicity⁴⁴.

When two-dimensional gels are used as a method of separating a qualitative subset of proteins, as opposed to comparing whole-cell preparations, or when immunological methods are used to highlight a subset of proteins, biologically relevant answers can be more readily obtained. For example, many secreted proteins can be identified by two-dimensional gel analysis of supernatants of cell lines and explants from tumour tissues⁴⁵. Several groups have probed two-dimensional gels of proteins from allergy-causing organisms using antibodies derived from allergic patients^{46,47}. Identification of the responsible allergen by mass spectrometry can be exploited in the rational design of preventive and therapeutic strategies.

We predict that protein expression analysis will be most useful in well-defined areas such as (1) analysis of samples that do not contain mRNA such as body fluids; (2) cases where the protein abundance does not correlate with the mRNA abundance; (3) cases where the critical changes involve post-translational modifications of proteins such as glycosylation or phosphorylation, rather than changes in protein abundance; (4) cases where an overview of the most abundant proteins in a specialized source is itself of importance; and (5) cases where two-dimensional gels allow a relatively comprehensive overview of a simple proteome such as that of a microbe.

Protein chips

In the protein chip approach, a variety of 'bait' proteins such as antibodies can be immobilized in an array format onto specially treated surfaces (Fig. 4). The surface is then probed with the sample of interest and only the proteins that bind to the relevant antibodies remain bound to the chip⁴⁸. Such an approach is essentially a large-scale version of enzyme-linked immunosorbent assays that are already used in clinical diagnostics. In one version, the protein chip is probed with fluorescently labelled proteins from two different cell states. Cell lysates are labelled by different fluorophores and mixed

Box 2

Mass spectrometric techniques in proteomics

MALDI and peptide-mass mapping

In this approach, a portion of the tryptic peptide mixture is analysed by MALDI mass spectrometry. Because trypsin cleaves the protein backbone at the amino acids arginine and lysine, the masses of tryptic peptides can be predicted theoretically for any entry in a protein sequence database. These predicted peptide masses are compared with those obtained experimentally by MALDI analysis. The protein can be identified correctly if there are a sufficient number of peptide matches with a protein in the database, resulting in a high score. High mass accuracy is critical for unambiguous identification and serves mainly to eliminate the false positives. MALDI identification by peptide-mass fingerprints requires that the full-length gene be present in the databases. Therefore, the success rate of this method will receive an additional boost with the availability of all predicted genes in sequence databases.

Electrospray and tandem mass spectrometry

There are two major mass spectrometric strategies that use electrospray ionization. In one of them the unseparated mixture of peptides is applied to a low-flow device called nanoelectrospray^{81,82}. The peptide mixture is then electrosprayed from a very fine needle (tip internal diameter of 1 μm) into the mass spectrometer. Individual peptides from the mixture are isolated in the first step and fragmented during the second step to sequence the peptide (hence tandem mass spectrometry). The fragments obtained by this method are derived from the N or C terminus of the protein and are designated 'b' or 'y' ions, respectively⁸³. The other strategy uses liquid chromatography for initial separation of peptides followed by sequencing as they elute into the electrospray ion source. This method can also be used without gel electrophoresis where a mixture of proteins is digested in solution and the 'scrambled' set of peptides are sequenced, ideally resulting in several peptide hits for each of the proteins that was initially present in the mixture^{54,84}. A great deal of data can be obtained from a single run in an automated fashion.

Using mass spectrometry data to search databases

In tandem mass spectrometry, peptides are fragmented by collision with gas molecules in the mass spectrometer. Spacing of these fragments by the molecular mass of one amino acid reveals the identity and location of that amino acid in the peptide. Only two such amino acids, combined with the knowledge of their location in the peptide — a 'peptide sequence tag' — is sufficient to locate the peptide in large sequence databases (Fig. 1c)^{85,86}. Alternatively, the theoretical fragmentation spectra of all possible peptides can be compared with the experimental spectrum to find the sequence that most likely gave rise to it^{87,88}. As a result, more complex mixtures of proteins can be analysed and the corresponding peptides found in EST databases or even directly in genomic databases. Routine sensitivities achieved by many laboratories are in the low picomole range (50–100 ng for most proteins), but specialized laboratories achieve higher sensitivities down to the low femtomole range of protein applied to the gel. The overall sensitivity of detection is determined mainly by the protein preparation methods as the mass spectrometer itself is capable of detecting sub-femtomole amounts of peptides under optimized conditions.

such that the colour acts as a readout for the change in abundance of the protein bound to the antibody. This system depends on reasonably specific and well-characterized antibodies and a number of technical problems would still need to be overcome. However, once developed it could provide convenient proteome analysis. In

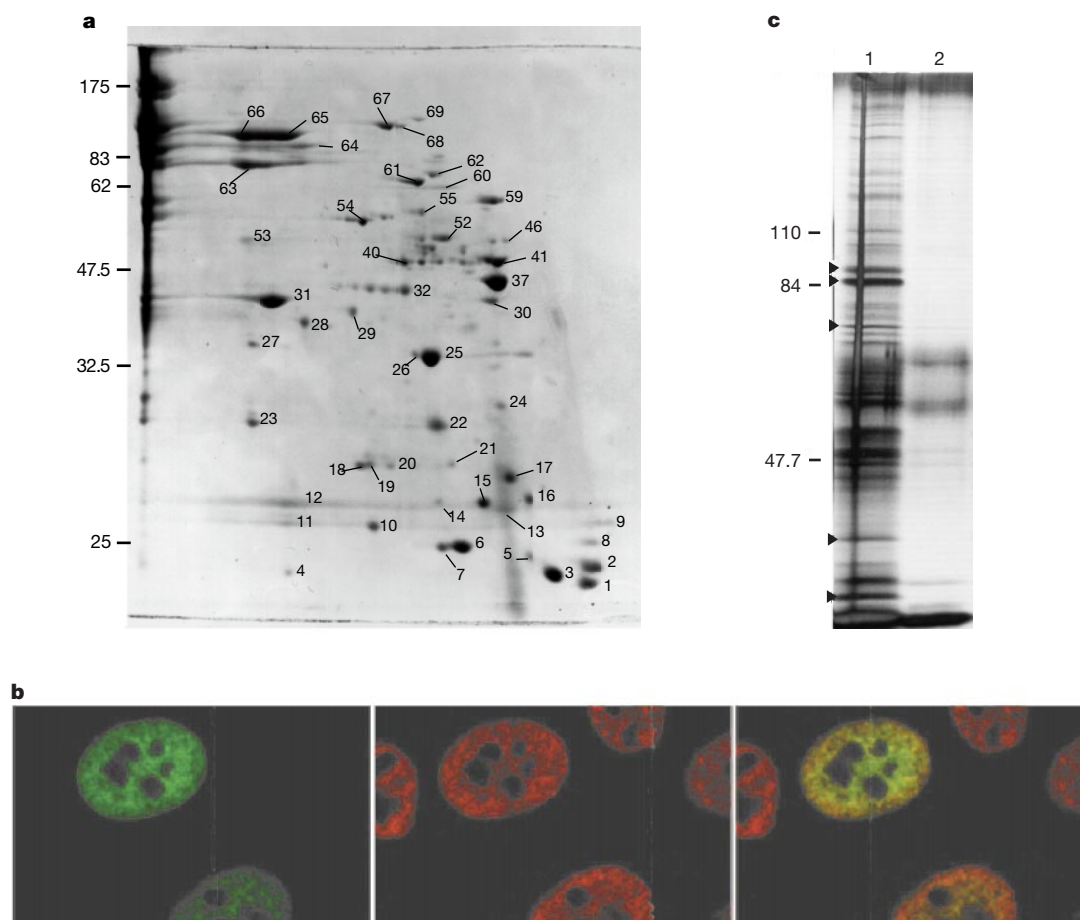


Figure 6 Characterization of the multi-protein spliceosome complex. **a**, A two-dimensional gel of spliceosome-associated factors. **b**, Expression of a green fluorescent protein (GFP)-tagged version of a protein, SPF45 (spot S28), identified from the gel shown in panel **a**. HeLa cells were transiently transfected with a plasmid encoding SPF45, which was tagged with GFP at its N terminus. The green fluorescence observed is due to localization of the GFP-tagged protein to the nucleus. Immunofluorescence using an antibody against a known nuclear protein, U1-specific snRNP protein or U1 (red signal), shows similar sub-nuclear localization as shown by the overlay (yellow signal). **c**, The strategy shown in Fig. 5 was used to isolate molecules interacting with S14. A one-dimensional gel showing proteins eluted from GST beads alone or GST-S14 is shown. The gel was silver stained and the bands indicated by arrowheads were excised and identified by mass spectrometry. These were again found to be proteins in the spliceosome complex, confirming the presence of S14 in the complex and providing insight to S14's role. (Fig. 6c courtesy of A. King.)

other modifications, peptides, protein fragments or proteins may also be immobilized onto chips and samples (for example, phage library or patient serum) applied onto the chip followed by detection of binding. One approach using protein chips couples the above techniques with a direct MALDI readout of the bound material^{49,50}.

Quantification by mass spectrometry

In addition to the above methods, differential-display proteomics can also be done using limited or no protein separation followed by mass spectrometric quantification. Because the intensity of a peptide peak in the mass spectrum cannot be predicted, quantification is achieved by labelling one of the two states by stable isotopes. Such methods have been used traditionally in mass spectrometry of small molecules but have only recently been applied to proteomics. Microbes can, for example, be grown in one state in normal medium and in another state in medium containing only N^{15} instead of N^{14} . Protein preparations from the two states are then mixed, separated and analysed by mass spectrometry. Two versions of any peptide can now be detected where one is greater in mass by its number of nitrogen atoms and the ratio of peak heights accurately quantifies the relative amounts of the corresponding proteins. As an alternative, Aebersold and colleagues introduced an isotopic non-radioactive label on cysteines after cell lysis before quantifying the samples by

mass spectrometry⁵¹. This strategy enables quantification of peptides from the most abundant components of very crude protein mixtures without gel electrophoresis.

Protein-protein interactions

A key question about a protein, in addition to when and where it is expressed, is with which other proteins does it interact. Interaction partners are an immediate lead into biological function and can potentially be exploited for therapeutic purposes. Creation of a protein-protein interaction map of the cell would be of immense value to understanding the biology of the cell.

Purification of protein complexes

Proteomics can make a key contribution to the study of protein-protein interactions⁵²⁻⁵⁵. An attractive way to study protein-protein interactions is to purify the entire multi-protein complex by affinity-based methods. This can be achieved in a variety of ways such as by using glutathione S-transferase (GST)-fusion proteins, antibodies, peptides, DNA, RNA or a small molecule binding specifically to a cellular target. One of the generic ways of identifying the interaction partners of a new protein is to tag it with an epitope. This protein can then be overexpressed in cells and — together with its interaction partners — immunoprecipitated by an antibody against the epitope.

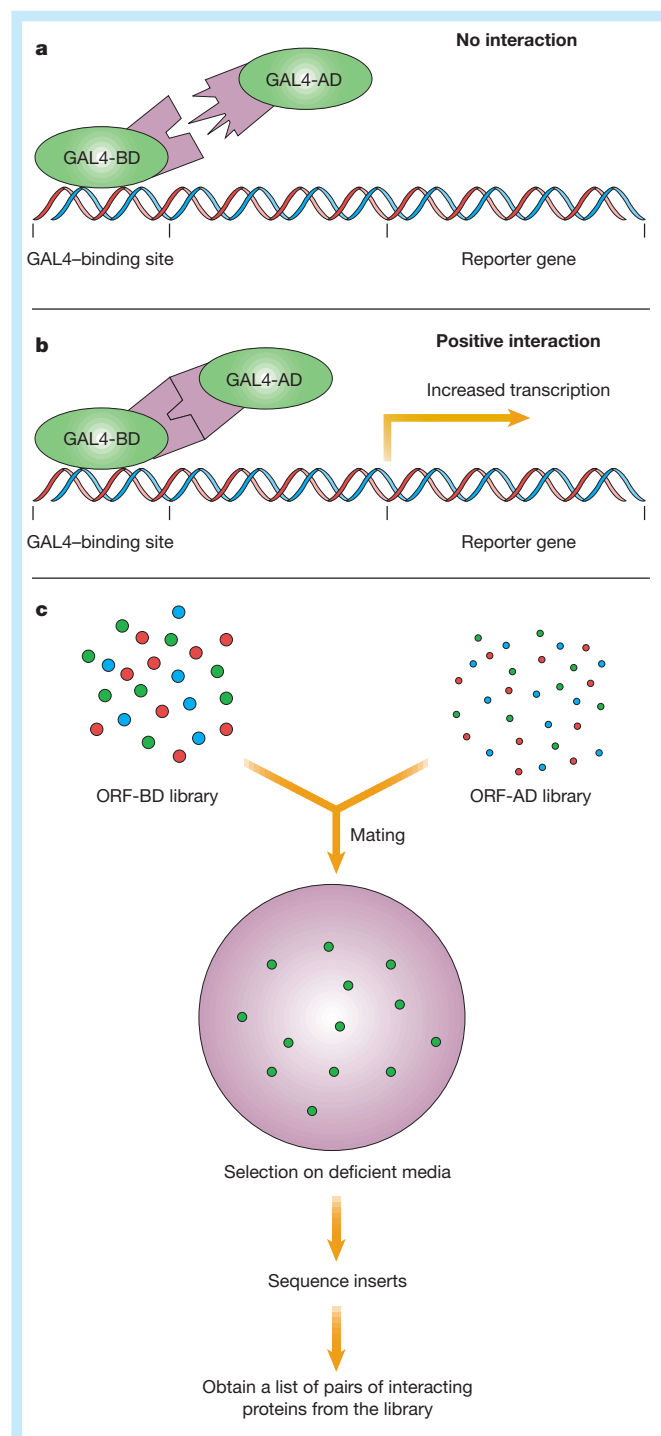


Figure 7 The yeast two-hybrid system. **a**, Different ORFs are expressed as fusion proteins to either the GAL4 DNA-binding domain (GAL4-BD) or its activation domain (GAL4-AD). If the proteins encoded by the ORFs do not interact with each other, the fusion proteins are not brought into close proximity and there is no activation of transcription of the reporter gene containing the upstream GAL4-binding sites. **b**, If the ORFs encode proteins that interact with each other, the fusion proteins are assembled at the GAL4-binding site of the reporter gene, which leads to activation of transcription. **c**, Library-based yeast two-hybrid screening method. In this strategy, two different yeast strains containing two different cDNA libraries are prepared. In one case, the ORFs are expressed as GAL4-BD fusions and in the other case, they are expressed as GAL4-AD fusions. The two yeast strains are then mated and diploids selected on deficient media. Thus, only the yeast cells expressing interacting proteins survive. The inserts from both the plasmids are then sequenced to obtain a pair of interacting genes.

This requires only the full-length complementary DNA clone of the gene and no time is spent in generating a precipitating antibody against the gene of interest. Because full-length cDNAs may soon be available for most human genes⁵⁶, large-scale interaction studies will become possible. Making fusion proteins such as GST-fusions is another generic way to obtain interaction partners (Fig. 5). The multi-protein complex associates with the 'bait', which is immobilized on a solid support. After washing away the proteins that interact nonspecifically, the protein complex is eluted, separated by gel electrophoresis and analysed by mass spectrometry. Thus, in a single experiment, the components of an entire multi-protein complex can be identified. As an example, the human spliceosome has been purified using biotinylated RNA as the 'bait' on which the complex assembled⁵⁷. Its protein components were then displayed by two-dimensional gel electrophoresis (Fig. 6a). From a single two-dimensional gel, 19 new factors were obtained (mostly in EST databases) and several of them were cloned and analysed further. Co-localization using immunofluorescence of the new protein with other members of the complex served to establish that they are bona fide members of the complex (Fig. 6b). Several of the new factors identified from this study were cloned and GST-fusion proteins generated. Using the strategy shown in Fig. 5, one of these proteins, designated S14, precipitated a subset of the spliceosome proteins (Fig. 6c), which, together with other experiments and bioinformatics analysis of the sequence, indicated a function of this protein. Many protein complexes have now been characterized using the strategy outlined above. Some of these complexes include the yeast Arp2/3 complex⁵⁸, proteins found in the yeast nuclear-pore complex⁵⁹ and proteins bound to the chaperonin GroEL⁶⁰.

These studies provide insight into mechanisms and open up new lines of investigations. Because no assumptions are made about the complex, unsuspected connections between cellular processes routinely emerge. For example, a study of profilin-I and -II binding proteins in mouse brain resulted in the discovery of two sets of proteins, one consisted of signalling molecules that regulate actin cytoskeleton and the other was involved in endocytosis. This indicated a link between signal transduction pathways and microfilament assembly involving profilin⁶¹.

Once members of a multi-protein complex have been identified by mass spectrometry, their function is studied by pertinent assays. At this stage, proteomics can be used in an iterative fashion to define either direct interaction partners of a new protein in the complex and/or to connect to other complexes in the cell⁶².

The success of the above-mentioned strategies relies on sufficient affinity of the protein complex to the bait and on optimized conditions for purification steps. For example, use of a double-tagging strategy improves complex recovery and reduces nonspecific protein binding⁶³. Lower-affinity interactions can potentially be captured by chemically crosslinking the protein complex before affinity purification because it relies on spatial proximity rather than affinity. Crosslinking can also help in elucidating the topological structure of a protein complex by the determination of nearest neighbours⁶⁴.

Components of specific organelles have also begun to be analysed. The yeast Golgi apparatus has been catalogued and the components of the chloroplast of garden pea have been similarly investigated to identify proteins involved in the processing, targeting, insertion and assembly of photosynthetic complexes^{65,66}. The interchromatin granules have been examined by the analysis of the crude peptide mixture obtained after digestion in solution of the entire sample⁶⁷.

Yeast two-hybrid system

The yeast two-hybrid system has emerged as a powerful tool to study protein-protein interactions⁶⁸. It is a genetic method based on the modular structure of transcription factors wherein close proximity of the DNA-binding domain to the activation domain induces increased transcription of a set of genes. The yeast hybrid system uses ORFs fused to the DNA-binding or -activation domain of GAL4 such that increased transcription of a reporter gene results when the

proteins encoded by two ORFs interact in the nucleus of the yeast cell (Fig. 7a, b). One of the main consequences of this is that once a positive interaction is detected, the ORF is identified simply by sequencing the relevant clones. For these reasons it is a generic method that is simple and amenable to high-throughput screening of protein–protein interactions.

On a large scale, this strategy has been used in two formats. In the array method, yeast clones containing ORFs as fusions to DNA or activation domains are arrayed onto a grid and the ORFs to be tested (as reciprocal fusions) are screened against the entire grid to identify interacting clones (Fig. 4). In the library screening method, one set of ORFs are first pooled to generate a library and then the reciprocal ORF–fusions are mated with the library one by one or several at a time (Fig. 7c).

Such analyses on a genome-wide scale have already been reported in *Saccharomyces cerevisiae* and to a more limited extent in *Caenorhabditis elegans*^{69–71}. In yeast, the array method was performed on 192 ORFs and the library screening method for 87% of the yeast genome. Together, this experiment resulted in 957 putative interactions⁷⁰. Another group analysed the results of 10% of an exhaustive library screen in yeast, resulting in 183 putative interactions⁷¹. The vast majority of the interactions found in these two large-scale studies were new. Several of these interactions seem plausible based on previous genetic or biochemical studies, whereas the relevance of most others cannot easily be determined. Therefore, such studies provide only potential interactions that have to be confirmed or eliminated by further biological experimentation. The main advantage of these methods is that they can be performed with a high throughput and in an automated manner. A recently described modification of the yeast two-hybrid method, termed ‘reverse’ two hybrid, can be used for identification of compounds and peptides that disrupt protein–protein interactions⁷². This can lead to development of drugs that have activities *in vivo* as opposed to drug screens that are conventionally done *in vitro*.

Phage display

Phage display is a method where bacteriophage particles are made to express either a peptide or protein of interest fused to a capsid or coat protein. It can be used to screen for peptide epitopes, peptide ligands, enzyme substrates or single-chain antibody fragments. Although combinatorial peptide libraries have generally been used in most phage display-based studies, more informative large-scale protein interaction studies can now be done if the products of cDNA libraries are displayed on phage particles. Any ‘bait’ protein can then be immobilized to capture phage particles displaying interacting proteins. This method is similar to the yeast two-hybrid system in that it is simple and can be performed with high throughput. Depending on the particular class of proteins being studied (such as cytoplasmic versus cell surface proteins), this method may be superior or inferior to the two-hybrid system because the interactions take place in solution as opposed to the nucleus of the yeast cell. Furthermore, this method is applicable in principle to transcription factors, which are not amenable to the yeast two-hybrid system. Methods have recently been optimized to display cDNA libraries on phages to isolate signalling molecules in the EGF-receptor signalling pathway as well as to identify antigens that react with certain antibodies^{73,74}.

Conclusions

Proteomics provides a powerful set of tools for the large-scale study of gene function directly at the protein level. In particular, the mass spectrometric study of gel-separated proteins is leading to a renaissance in biochemical approaches to protein function. Protein characterization will continue to improve in throughput, sensitivity and completeness. Post-translational modifications cannot currently be studied at high throughput but certain categories such as phosphorylation are beginning to be amenable to generic approaches. We predict that proteomics will move away from the monitoring of protein expression using two-dimensional gels. Mass

spectrometry-based methods that use affinity purification followed by only one-dimensional electrophoresis will continue to gain in importance. In the near future, proteomics will provide a wealth of protein–protein interaction data, which will probably be its most important and immediate impact on biological science. Because proteins are one step closer to function than are genes, these studies frequently lead directly to biological discoveries or hypotheses. The ready availability of many human genes as full-length clones is itself an extremely important extension of the genome projects that will make possible several proteomic strategies. Assays to determine protein function using purified proteins will be automated and performed in miniaturized grid formats in parallel for thousands of proteins. Finally, advances in genomics will directly fuel large-scale protein assays that use genetics as a readout, such as the two-hybrid screen. □

- Wilkins, M. R., Williams, K. L., Apple, R. D. & Hochstrasser, D. F. *Proteome Research: New Frontiers in Functional Genomics* 1–243 (Springer, Berlin, 1997).
- Wilkins, M. R. *et al.* From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *BioTechnology* **14**, 61–65 (1996).
- Celis, J. *et al.* Human 2-D PAGE databases for proteome analysis in health and disease: <http://biobase.dk/cgi-bin/celis>. *FEBS Lett.* **398**, 129–134 (1996).
- Anderson, N. G. & Anderson, N. L. Twenty years of two-dimensional electrophoresis: past, present and future. *Electrophoresis* **17**, 443–453 (1996).
- O’Farrell, P. H. High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.* **250**, 4007–4021 (1975).
- Burley, S. K. *et al.* Structural genomics: beyond the human genome project. *Nature Genet.* **23**, 151–157 (1999).
- Krogh, A. in *Guide to Human Genome Computing* (ed. Bishop, M. J.) 261–274 (Academic, San Diego, 1998).
- Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999).
- Claverie, J. M. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6**, 1735–1744 (1997).
- Pandey, A. & Lewitter, F. Nucleotide sequence databases: a gold mine for biologists. *Trends Biochem. Sci.* **24**, 276–280 (1999).
- Brenner, S. E. Errors in genome annotation. *Trends Genet.* **15**, 132–133 (1999).
- Gygi, S., Rochon, Y., Franz, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
- Futcher, B. *et al.* A sampling of the yeast proteome. *Mol. Cell. Biol.* **19**, 7357–7368 (1999).
- Pandey, A., Andersen, J. S., & Mann, M. Use of mass spectrometry to study signaling pathways. *Science’s STKE* (in the press).
- Jensen, O. N., Mortensen, P., Vorm, O. & Mann, M. Automation of matrix assisted laser desorption/ionization mass spectrometry using fuzzy logic feedback control. *Anal. Chem.* **69**, 1706–1714 (1997).
- Berndt, P., Hohmann, U. & Langen, H. Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. *Electrophoresis* **20**, 3521–3526 (1999).
- Shevchenko, A. *et al.* Linking genome and proteome by mass spectrometry: large scale identification of yeast proteins from two dimensional gels. *Proc. Natl Acad. Sci. USA* **93**, 14440–14445 (1996).
- Shevchenko, A. *et al.* MALDI quadrupole time-of-flight mass spectrometry: powerful tool for proteomic research. *Anal. Chem.* **72**, 2132–2141 (2000).
- Zhang, B., Liu, H., Karger, B. L. & Foret, F. Microfabricated devices for capillary electrophoresis-electrospray mass spectrometry. *Anal. Chem.* **71**, 3258–3264 (1999).
- Figeys, D., Gygi, S. P., McKinnon, G. & Aebersold, R. An integrated microfluidics-tandem mass spectrometry system for automated protein analysis. *Anal. Chem.* **70**, 3728–3734 (1998).
- Li, J. *et al.* Integration of microfabricated devices to capillary electrophoresis—electrospray mass spectrometry using a low dead volume connection: application to rapid analyses of proteolytic digests. *Anal. Chem.* **71**, 3036–3045 (1999).
- Eckerkorn, C. *et al.* Mass spectrometric analysis of blotted proteins after gel electrophoresis separation by matrix-assisted laser desorption/ionization. *Electrophoresis* **13**, 664–665 (1992).
- Strupat, K. *et al.* Matrix-assisted laser desorption/ionization mass spectrometry of proteins electroblotted after polyacrylamide gel electrophoresis. *Anal. Chem.* **66**, 464–470 (1994).
- Bienvenut, W. V. *et al.* Toward a clinical molecular scanner for proteome research: parallel protein chemical processing before and during western blot. *Anal. Chem.* **71**, 4800–4807 (1999).
- Binz, P. A. *et al.* A molecular scanner to automate proteomic research and to display proteome images. *Anal. Chem.* **71**, 4981–4988 (1999).
- Jensen, P. K. *et al.* Probing proteomes using capillary isoelectric focusing-electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **71**, 2076–2084 (1999).
- Mortz, E. *et al.* Sequence tag identification of intact proteins by matching tandem mass spectral data against sequence data bases. *Proc. Natl. Acad. Sci. USA* **93**, 8264–8267 (1996).
- Li, W., Hendrickson, C. L., Emmett, M. R. & Marshall, A. G. Identification of intact proteins in mixtures by alternated capillary liquid chromatography electrospray ionization and LC ESI infrared multiphoton dissociation Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **71**, 4397–4402 (1999).
- Nuwaysir, L. & Stults, J. T. ESI mass spectrometry of phosphopeptides isolated by on-line immobilized metal affinity chromatography. *J. Am. Soc. Mass Spectrom.* **4**, 662–669 (1993).
- Betts, J. C., Blackstock, W. P., Ward, M. A. & Anderton, B. H. Identification of phosphorylation sites on neurofilament proteins by nanoelectrospray mass spectrometry. *J. Biol. Chem.* **272**, 12922–12927 (1997).
- Neubauer, G. & Mann, M. Mapping of phosphorylation sites of gel-isolated proteins by nanoelectrospray

- tandem mass spectrometry: potentials and limitations. *Anal. Chem.* **71**, 235–242 (1999).
33. Zhang, X. *et al.* Identification of phosphorylation sites in proteins separated by polyacrylamide gel electrophoresis. *Anal. Chem.* **70**, 2050–2059 (1998).
34. Cortez, D., Wang, Y., Qin, J. & Elledge, S. J. Requirement of ATM-dependent phosphorylation of bcr1 in the DNA damage response to double-strand breaks. *Science* **286**, 1162–1166 (1999).
35. Soskic, V. *et al.* Functional proteomics analysis of signal transduction pathways of the platelet-derived growth factor beta receptor. *Biochemistry* **38**, 1757–1764 (1999).
36. Pandey, A. *et al.* Analysis of receptor signaling pathways by mass spectrometry: identification of Vav-2 as a substrate of the epidermal and platelet-derived growth factor receptors. *Proc. Natl Acad. Sci. USA* **97**, 179–184 (2000).
37. DeRisi, J., Iyer, V. R. & Brown, O. P. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
38. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
39. Roberts, C. J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
40. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
41. Ostergaard, M., Wolf, H., Orntoft, T. F. & Celis, J. E. Psoriasin (S100A7): a putative urinary marker for the follow-up of patients with bladder squamous cell carcinomas. *Electrophoresis* **20**, 349–354 (1999).
42. Page, M. J. *et al.* Proteomic definition of normal human luminal and myoepithelial breast cells purified from reduction mamoplasties. *Proc. Natl Acad. Sci. USA* **96**, 12589–12594 (1999).
43. Gaus, C. *et al.* Analysis of the mouse proteome. (I) Brain proteins: separation by two-dimensional electrophoresis and identification by mass spectrometry and genetic variation. *Electrophoresis* **20**, 575–600 (1999).
44. Aicher, L. *et al.* New insights into cyclosporine A nephrotoxicity by proteome analysis. *Electrophoresis* **19**, 1998–2003 (1998).
45. Celis, J. E. *et al.* A comprehensive protein resource for the study of bladder cancer: <http://biobase.dk/cgi-bin/celis>. *Electrophoresis* **20**, 300–309 (1999).
46. Breitenbach, M. *et al.* Biological and immunological importance of Bet v 1 isoforms. *Adv. Exp. Med. Biol.* **409**, 117–126 (1996).
47. Sander, I. *et al.* Allergy to aspergillus-derived enzymes in the baking industry: identification of beta-xylosidase from aspergillus niger as a new allergen (Asp n 14). *J. Allergy Clin. Immunol.* **102**, 256–264 (1998).
48. Lueking, A., Horn, M., Eickhoff, H., Lehrach, H. & Walter, G. Protein microarrays for gene expression and antibody screening. *Anal. Biochem.* **270**, 103–111 (1999).
49. Davies, H., Lomas, L. & Austen, B. Profiling of amyloid beta peptide variants using SELDI Protein Chip arrays. *Biotechniques* **27**, 1258–1261 (1999).
50. Nelson, R. W. The use of bioreactive probes in protein characterization. *Mass Spectrom. Rev.* **16**, 353–376 (1997).
51. Gygi, S. P. *et al.* Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.* **17**, 994–999 (1999).
52. Neubauer, G. *et al.* Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl Acad. Sci. USA* **94**, 385–390 (1997).
53. Lamond, A. I. & Mann, M. Cell biology and the genome projects—a concerted strategy for characterizing multi-protein complexes using mass spectrometry. *Trends Cell Biol.* **7**, 139–142 (1997).
54. Link, A. J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnol.* **17**, 676–682 (1999).
55. Blackstock, W. P. & Weir, M. P. Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* **17**, 121–127 (1999).
56. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. The mammalian gene collection. *Science* **286**, 455–457 (1999).
57. Neubauer, G. *et al.* Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nature Genet.* **20**, 46–50 (1998).
58. Winter, D., Podtelejnikov, A. V., Mann, M. & Li, R. The complex containing actin-related proteins Arp2 and Arp3 is required for the motility and integrity of yeast actin patches. *Curr. Biol.* **7**, 519–529 (1997).
59. Rout, M. P. *et al.* The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell. Biol.* **148**, 635–651 (2000).
60. Houry, W. A. *et al.* Identification of in vivo substrates of the chaperonin GroEL. *Nature* **402**, 147–154 (1999).
61. Witke, W. *et al.* In mouse brain profilin I and profilin II associate with regulators of the endocytic pathway and actin assembly. *EMBO J.* **17**, 967–976 (1998).
62. Shevchenko, A. & Mann, M. in *Mass Spectrometry in Biology and Medicine* (eds Burlingame, A., Carr, C. A. & Baldwin, M. A.) 237–269 (Humana, Totowa, 1999).
63. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17**, 1030–1032 (1999).
64. Rappsilber, J., Siniosoglou, S., Hurt, E. C. & Mann, M. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal. Chem.* **72**, 267–275 (2000).
65. Rowley, A. *et al.* Applications of protein mass spectrometry in cell biology. *Methods* **20**, 383–397 (2000).
66. Peltier, J. B. *et al.* Proteomics of the chloroplast. Systematic identification and targeting analysis of luminal and peripheral thylakoid proteins. *Plant Cell* **12**, 319–342 (2000).
67. Mintz, P. J. *et al.* Purification and biochemical characterization of interchromatin granule clusters. *EMBO J.* **18**, 4308–4320 (1999).
68. Fields, S. & Song, O. K. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246 (1989).
69. Walhout, A. J. *et al.* Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* **287**, 116–122 (2000).
70. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
71. Ito, T. *et al.* Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA* **97**, 1143–1147 (2000).
72. Vidal, M. & Endoh, H. Prospects for drug screening using the reverse two-hybrid system. *Trends Biotechnol.* **17**, 374–381 (1999).
73. Zozulya, S. *et al.* Mapping signal transduction pathways by phage display. *Nature Biotechnol.* **17**, 1193–1198 (1999).
74. Hufton, S. E. *et al.* Phage display of cDNA repertoires: the pVI display system and its applications for the selection of immunogenic ligands. *J. Immunol. Methods* **231**, 39–51 (1999).
75. Martnez, M. R. *et al.* A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286**, 1153–1155 (1999).
76. Zambrowicz, B. P. *et al.* Disruption and sequence identification of 2,000 genes in mouse embryonic stem cells. *Nature* **392**, 608–611 (1998).
77. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
78. Winzler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
79. Mattheakis, L. C., Bhatt, R. R. & Dower, W. J. An in vitro polysome display system for identifying ligands from very large peptide libraries. *Proc. Natl Acad. Sci. USA* **91**, 9022–9026 (1994).
80. Roberts, R. W. & Szostak, J. W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl Acad. Sci. USA* **94**, 12297–12302 (1997).
81. Wilm, M. & Mann, M. Analytical properties of the nanoelectrospray ion source. *Anal. Chem.* **68**, 1–8 (1996).
82. Wilm, M. *et al.* Femtomole sequencing of proteins from polyacrylamide gels by nano electrospray mass spectrometry. *Nature* **379**, 466–469 (1996).
83. Roepstorff, P. & Fohlman, J. Proposed nomenclature for sequence ions. *Biomed. Mass Spectrom.* **11**, 601 (1984).
84. Yates, J. R. Mass spectrometry: From genomics to proteomics. *Trends Genet.* **16**, 5–8 (2000).
85. Mann, M. & Wilm, M. S. Error tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **66**, 4390–4399 (1994).
86. Mann, M. A shortcut to interesting human genes: peptide sequence tags, ESTs and computers. *Trends Biochem. Sci.* **21**, 494–495 (1996).
87. Eng, J. K., McCormack, A. L. & J. R. Yates, I. An approach to correlate MS/MS data to amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
88. Yates, J. R. III Database searching using mass spectrometry data. *Electrophoresis* **19**, 893–900 (1998).

Acknowledgements

We thank B. Blagoev and M. Fernandez for their expert assistance with cell culture and immunoprecipitation experiments. We thank all other members of the Protein Interaction Laboratory for valuable discussions and comments on the manuscript and A. King, Protana A/S, for obtaining the data for the new spliceosomal protein. O. N. Jensen and A. Stensballe are acknowledged for their contributions in the analysis of phosphopeptides. A.P. was supported by the Howard Temin Award from the National Cancer Institute. This work was funded in part by a grant from the Danish National Research Foundation to M.M.'s laboratory (www.pil.sdu.dk) at the Center for Experimental Bioinformatics (CEBI).

Searching for genetic determinants in the new millennium

Neil J. Risch

Department of Genetics, Stanford University School of Medicine, Stanford, California 94305-5120, USA

Human genetics is now at a critical juncture. The molecular methods used successfully to identify the genes underlying rare mendelian syndromes are failing to find the numerous genes causing more common, familial, non-mendelian diseases. With the human genome sequence nearing completion, new opportunities are being presented for unravelling the complex genetic basis of non-mendelian disorders based on large-scale genome-wide studies. Considerable debate has arisen regarding the best approach to take. In this review I discuss these issues, together with suggestions for optimal post-genome strategies.

It is now 135 years since the Bohemian monk Gregor Mendel published the results of his breeding experiments on the garden pea, which initiated the modern era of the study of genetics. In Mendel's time, the abounding theory of heredity postulated a 'blending' of the inherited contributions from the two parents. Mendel's work clearly showed that such blending did not occur, and led to his conclusion of particulate inheritance (the 'gene') and rules of segregation. The relevance of Mendel's work for human traits was first delineated around the turn of the century by Garrod, who reasoned correctly that similar types of transmission rules explained the 'inborn errors of metabolism' typically caused by enzyme deficiencies. At the same time, however, there was another school of thought, primarily emanating from statisticians such as Francis Galton and his student, Karl Pearson. They observed family resemblance for a variety of traits such as anthropometric features and intellectual achievement but they could not discern patterns of inheritance in families that were consistent with mendelian laws. Rather, a 'blending'-type theory seemed more apt, as children's phenotypes tended to be, on average, midway between the parents, with some variability. The resolution of this dilemma did not appear until 1918, when Ronald Fisher published his seminal paper describing 'polygenic' inheritance. Fisher reconciled the two conflicting schools by recognizing that the critical difference lay in the genetic basis for the variation in the trait being studied.

For the traits Mendel studied, the observed variation was due to a simple difference at a single gene (or locus). On the other hand, for the traits studied by the biometrical school, individual differences were not attributable to different alleles at a single locus. Rather, many different genes, each with allelic variations, contributed to the total observed variability in a trait, with no particular gene having a singly large effect. Thus, an individual phenotype results from the sum total of the effects of all the numerous contributing loci. Furthermore, application of the central limit theorem from statistics implicates a continuous normal distribution in the population for such a trait, similar to what is observed. Thus, the lack of mendelian inheritance patterns for numerous human traits did not require the deconstruction of Mendel's theory, but rather an extension of it to a more complex scenario that related genes to phenotype. It is clear that Mendel's success hinged

entirely on his selection of single-gene traits, for otherwise the simple rules of inheritance would not have revealed themselves.

The past two decades has witnessed an explosion in both molecular and computational technology, which has enabled the identification of genes for a number of inherited human disorders. These successes have been restricted largely to simple mendelian cases, which are by their nature rare, and although important to the individuals who carry these genes, of limited significance in terms of public health. The promise of the same technology solving the problem of more frequent, non-mendelian familial disorders has largely been unfulfilled. At the same time, at this turn of the millennium, we now find ourselves at the threshold of having the entire human DNA sequence on hand (or at least in silica). It is therefore timely to consider how this new information can best be used in future gene-finding studies, and prospects for success.

The genetic basis of human traits and disease

Critical to the discussion of what approaches are best suited to unravel the genetic basis of traits or disease in the new millennium is a working model of what that basis is likely to entail. So far, we still have a view that primarily reflects the Mendelist-biometricist dialogue of nearly a century ago. Most human disorders that have been genetically characterized are mendelian, essentially because the extant molecular tools have enabled the identification of these genes by positional cloning (described later), a procedure now described as 'routine'. By contrast, those disorders or traits for which such approaches have failed are depicted as 'polygenic', multifactorial or 'complex'. Often unwilling to cede to a notion of 'infinite' genetic complexity, geneticists refer to these cases as 'oligogenic' or 'multigenic', implicating a tractable degree of complexity.

If one considers that there are estimated to be approximately 100,000 functional genes in humans and functional variation may exist in any of them, the problem becomes apparent. If the genetic variation that contributes to a trait is due to myriad genes, each of modest effect, the task of identifying those individual contributors becomes monumental. The fact is, however, that gene effects typically come in different sizes, even when there are many of them—at least, this has been the lesson from a lengthy history of model systems. There are several measures of gene effects used by geneticists (Box 1). Many human traits, especially disease

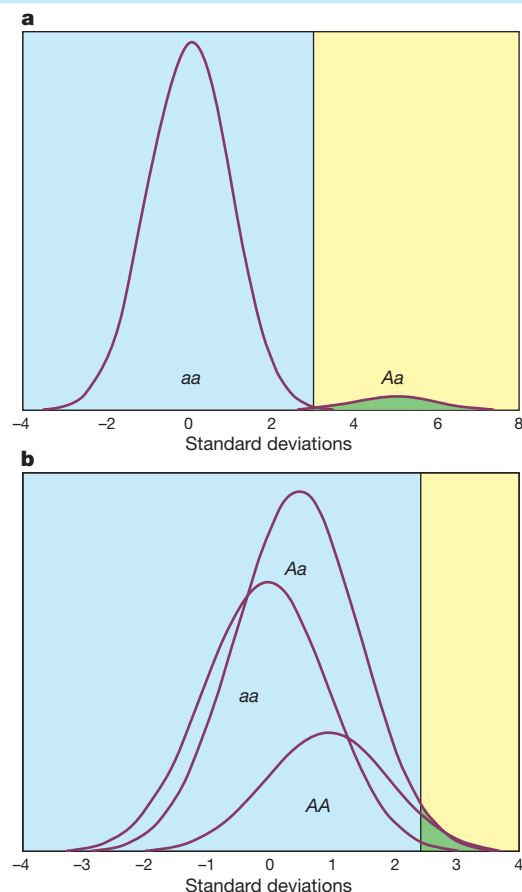


Figure 1 Examples of mendelian and non-mendelian inheritance using a gaussian model. Both loci have the same heritability $H^2_L = 12\%$. **a**, Dominant mendelian locus with allele frequency $p = 0.00275$ and displacement $t = 5$ s.d. Disease occurs above the threshold of 3 s.d. Disease risk for heterozygotes (Aa) is 98% and for homozygotes (aa) it is 0.13%. The population prevalence $K = 0.67\%$. **b**, Non-mendelian additive locus with allele frequency $p = 0.40$ and displacement $t = 0.5$ s.d. for each A allele (or total displacement $t = 1$). Disease occurs above the threshold of 2.5 s.d. Disease risk for high-risk homozygotes (AA) is 6.7%, for heterozygotes (Aa) it is 2.3% and for low-risk homozygotes (aa) it is 0.62%. The population disease prevalence $K = 2.4\%$. Even though the locus is additive on the liability scale, the disease risks are non-additive.

outcomes, show family recurrence patterns that are strongly suggestive of interactions between genes or epistasis, implying the existence of multiple, interacting loci.

Finding genes — a historical perspective

Before the early 1980s, genetic risk factors for a disease or trait could be identified only through direct analysis of candidate genes, usually through association studies. Starting soon after their discovery, blood-group systems such as ABO, MN and Rh were tested directly against an array of human diseases, typically with little replicability. However, after the study of tens of thousands of subjects, it seems that ABO shows consistent, but weak, association with a number of traits involving the gastrointestinal tract¹.

Case-control studies

The approach often used for such studies is the case-control design, in which a difference in allele frequency is sought between affected individuals and unrelated unaffected controls. From an epidemiological perspective, a major limitation in this approach is the potential for confounding (that is, spurious association resulting

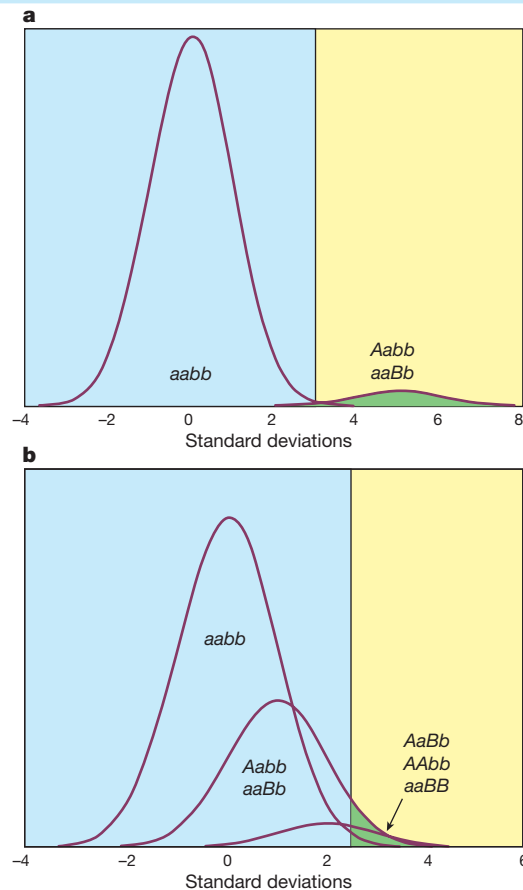


Figure 2 Examples of two-locus genetic models. **a**, Genetic heterogeneity with two rare dominant mendelian alleles (A and B) each with a frequency $p = 0.01$. The displacement t for each A and B allele is 5 s.d. Disease risk for each heterozygote is 98% whereas for normal homozygotes it is 0.13%. Other genotypes are extremely rare. Population disease prevalence $K = 4\%$. **b**, Additive non-mendelian model. The A and B allele each have frequency $p = 0.10$. Displacement is 1 s.d. for each A or B allele, or total displacement $t = 2$ for each locus. Disease occurs above a threshold of 2.5 s.d. Disease risk for genotype $aabb$ is 0.62%; for genotypes $Aabb$ and $aaBb$ it is 6.7%; for genotypes $AaBb$, $AABb$, $aaBB$ it is 31%; and for genotypes $AaBB$, $AaBB$ (rare, not shown) it is 69%. Population disease prevalence $K = 4\%$. Although the two loci are additive on the liability scale, the disease risks are non-additive and show both dominance and epistasis effects.

from correlation with the true risk factor) leading to artefactual as opposed to causal associations. In this case, the most likely source of confounding is ethnicity, whereby allele frequencies vary by ethnicity and cases and controls are not adequately matched in terms of ethnicity. Although most investigators would at least attempt coarse matching by major demographic groupings (such as race), substratification within racial groups can still lead to bias. This drawback of traditional case-control designs was recognized early on by Lionel Penrose, who recommended the use of unaffected sibs as controls². This paradigm, originally applied to ABO and duodenal ulcer³, has seen a resurgence in the past few years⁴⁻⁸. The disadvantage of this design is that sib controls are over-matched to the index cases, leading to a loss of power compared with a well-designed study involving unrelated controls⁷.

Conventional case-control gene-association studies have a long track record of false-positive results. The high false-positive rate has often been attributed to confounding due to stratification, although this has never been proven. It is more likely that the high false-positive rate results from a low prior probability that the few gene

Box 1

Measuring gene effects

The notion of numerous factors contributing to a trait, be they genetic or environmental, lends itself naturally to a gaussian (or normal) population distribution for the cumulative effect of those factors. This is the fundamental precept of biometrical genetics. Thus, for continuously measurable traits, it is simplest to characterize gene effects in the context of the gaussian distribution. If we consider a single locus L with two variant alleles A and a with population frequencies p and $q = 1 - p$ respectively, there are two important measures we can define⁴⁴. The first is termed displacement, denoted by t , which is the number of standard deviations difference between the mean values of the two homozygotes AA and aa (we assume, for simplicity, that the variance within genotype is the same for each genotype). There is an additional parameter, d , representing the mean value of heterozygotes Aa relative to the two homozygotes. Thus, a value of $d = 1$ corresponds to equal means for genotypes AA and Aa (that is, A is dominant), whereas $d = 0$ corresponds to equal means for genotypes Aa and aa (that is, A is recessive). A value of $d = 0.5$ corresponds to the heterozygotes being exactly intermediate between the two homozygotes, a situation often described as additive. The second important measure of gene effect is the population variance attributable to segregation of the gene. This is given by $V_G(L) = V_A(L) + V_D(L)$, where $V_A(L) = 2pq t^2(p(1-d) + qd)^2$ is the 'additive' genetic variance and $V_D(L) = p^2 q^2 (d - 0.5)^2$ is the 'dominance' genetic variance. The proportion of total variance attributable to locus A , which we denote h_L^2 , is then given by $V_G(L)/(1 + V_G(L))$, assuming the variance within genotype to be 1.0. It is important to note that h_L^2 is a function of both displacement t and the allele frequency p . Thus, a rare gene with large displacement (that is, mendelian) may contribute the same proportion to variance as a common gene with modest displacement (that is, non-mendelian; see Fig. 1). In addition, because h_L^2 is a function of p , its value can vary from one population to another when p varies, even when the displacement t is the same.

The gaussian model described above has a direct extension to discrete outcomes (for example, affected/unaffected). In this case, the quantitative genetic variable is latent (unobserved) and often described as the genetic 'liability'. Superimposed on the genetic-liability distribution is a risk function, so that risk of disease also increases continuously with liability. If one assumes a gaussian form for this risk function, the model can be characterized as a 'threshold' model, wherein total liability is defined as the sum of

genetic and non-genetic liabilities, and disease occurs when an individual's total liability exceeds a threshold T . As in the case of continuous outcomes, a single locus can influence the distribution of liability, where individual genotypes have different mean liabilities (see Fig. 1). The same measures used for quantitative outcomes can also be used here — namely displacement and proportion of variance explained, measured on the scale of liability.

Alternatively, one can conceptualize measures of gene effect on the scale of risk rather than liability. For example, classical measures from epidemiology, such as the relative risk, can be used to quantify the risk of disease for one genotype (say AA) compared to another (say aa), a concept termed the genotype relative risk or genotypic risk ratio (GRR)⁴⁵. The GRR is analogous to displacement, in that it measures the effect of a particular allele or genotype, independent of its frequency. Another useful, more complex measure is the sibling relative risk (λ_s) attributable to locus L (ratio of risk to sibs of an affected case to the population prevalence)⁴⁶. If the GRR for genotype AA is g_2 and that for genotype Aa is g_1 , and the frequency of allele A is p and $q = 1 - p$, λ_s can be calculated as $1 + (1/2V_A + 1/4V_D)/K^2$, where $K = p^2 g_2 + 2pq g_1 + q^2$, $V_A = 2pq(p(g_2 - g_1 + q(g_1 - g_0)))^2$ and $V_D = p^2 q^2 (g_2 + g_0 - 2g_1)^2$. Note that these formulas are analogous to the continuous case except that the displacement t is now replaced by the GRR g_2 .

The measures described above for effects of single loci need to be considered in the context of their genetic and/or environmental background. The gaussian model provides the simplest context whereby all other genetic factors are assumed to have small and additive effects, and the environment is also assumed to be gaussian and additive. In this case, in addition to the components of genetic variance $V_A(L)$ and $V_D(L)$ defined for locus L , we have the components of residual genetic variance $V_A(R)$ and $V_D(R)$, which are the additive and dominance variance summed across all other loci, with $V_G(R) = V_A(R) + V_D(R)$. The non-genetic component is assumed to have variance V_E .

From the perspective of biometrical genetics, epistasis refers to non-additive interactions between gene effects (much as dominance refers to non-additive effects between alleles at a single locus). Thus, the genetic variance underlying a trait can include sources of variance involving interactive effects among any number of loci, and these are termed epistatic variance components. Often, these are segregated into terms based on the number of loci involved in the interaction (for example, two, three or four loci)⁴⁴.

polymorphisms examined are in fact causally related to the disease outcomes studied. A case in point relates to another locus (or set of loci) for which the track record has been much better — the human leukocyte antigen (HLA) system on the short arm of chromosome 6 (chromosome 6p). Associations between specific HLA antigens and a variety of diseases (mostly autoimmune) have been reported and repeatedly confirmed — for example, with insulin-dependent diabetes mellitus, multiple sclerosis, rheumatoid arthritis, psoriasis, celiac disease, narcolepsy, haemochromatosis, and many others. The greater success rate in this case reflects the much higher prior probability of a causal relationship for this complex of loci than for other tested loci.

Linkage analysis and positional cloning

The situation of gene discovery in humans changed markedly two decades ago when it was recognized that variations in human DNA could be assayed directly and used as genetic markers in linkage studies⁹. The evolution of the field since then has been nothing short of dramatic. Before this time, human geneticists performing linkage studies to identify the chromosomal location of disease genes relied on only a handful of blood group and serum protein markers with

few successes. The identification of restriction-fragment length polymorphism (RFLP) markers⁹ and subsequently abundant highly polymorphic microsatellite (short tandemly repetitive DNA) loci^{10,11} has led to the mapping of myriad mendelian disease loci. Development of more efficient molecular tools, especially high-throughput DNA sequencing, has enabled the identification of disease loci and their mutations by a process characterized as positional cloning. Naturally occurring mutations are identified on the basis of their chromosomal location by taking advantage of the meiotic process of recombination as manifest in families segregating for the disease. Markers closest to the disease gene show the strongest correlation with disease patterns in families, and typically the tracking of recombination events can narrow the region harbouring a disease gene to between 100 and several thousand kilobases.

The remarkable success of positional cloning rests not simply on the advances observed in molecular technology. It also reflects the enormous power of linkage analysis when applied to mendelian phenotypes — that is, those characterized by a (near) one-to-one correspondence between genotypes at a single locus and the observed phenotype (a glossary of terms is presented in Box 3). In terms of

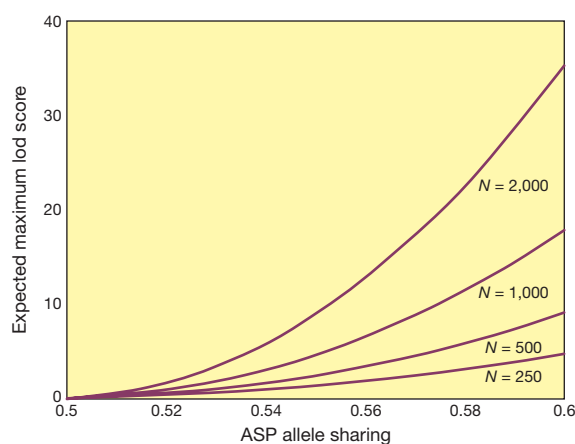


Figure 3 Range of number of ASPs required to detect linkage as a function of allele sharing.

biometrical genetics, these are loci with very high displacement (Fig. 1). The observed phenotype corresponds precisely to the underlying genotype with little if any misclassification. The robustness of linkage analysis applied to mendelian traits can be seen by its historic low false-positive rate¹² when the stringent lod-score threshold of 3 suggested by Morton¹³ is used (corresponding to a P value of 10^{-3} for a sequential test or 10^{-4} for a fixed sample-size test¹⁴). As I will discuss later, this conclusion is true only for the study of mendelian traits.

Genetic heterogeneity in mendelian disease

An important issue in the study of mendelian disease is the phenomenon of genetic heterogeneity, whereby distinct mutations at the same locus (allelic heterogeneity) or different loci (non-allelic heterogeneity) can cause the same, indistinguishable phenotype. Non-allelic genetic heterogeneity is a form of multi-locus model, wherein the predisposing alleles at each locus are typically rare and independently capable of producing disease. By contrast, common predisposing alleles often lead to epistasis or interaction effects among loci (Fig. 2). In linkage analysis, allelic heterogeneity does not cause a problem because all families (including those with different mutations) will show linkage to the same chromosomal region. In fact, allelic heterogeneity also provides the strongest evidence for a causal relationship between a cloned gene and disease phenotype. Statistically, it is extraordinarily unlikely to find several different mutations at the same locus in unrelated families with the same disease.

Non-allelic heterogeneity can cause a problem in linkage analysis, depending on its extent. In the extreme situation that any single gene accounts for a small proportion of segregating families, very large

families would be required to obtain robust linkage evidence, and positional cloning would still be difficult. But for mendelian disease this has rarely, if ever, been the case. More typically, when non-allelic heterogeneity exists, it involves only a few distinct loci; this degree of heterogeneity usually is not a serious impediment either to linkage analysis or positional cloning, essentially because the relationship between phenotype and genotype within families remains strong.

Another important issue relating to mutational heterogeneity is the population under study. For mendelian disease, endogamous population isolates with a limited number of founders tend to have less mutational heterogeneity and an increased frequency of founder effects, which makes them particularly useful in studies of positional cloning. When most affected individuals in a population carry a mutation derived from a single ancestor, they effectively create a single large extended pedigree, although most of the distant relationships are missing. Historic recombination events around the disease mutation can still be inferred, however, by examining the extent of DNA shared on present-day disease chromosomes. This approach, referred to as linkage disequilibrium analysis, has been highly effective in leading to the cloning of numerous disease genes.

The challenge of non-mendelian inheritance

As noted above, linkage analysis and positional cloning have had a remarkable track record in leading to the identification of the genes for many mendelian diseases, all within the time span of the past two decades. Several of these genes account for an uncommon subset of generally more common disorders such as breast cancer (BRCA-1 and -2), colon cancer (familial adenomatous polyposis (FAP) and hereditary non-polyposis colorectal cancer (HNPCC)), Alzheimer's disease (β -amyloid precursor protein (APP) and presenilin-1 and -2) and diabetes (maturity-onset diabetes of youth (MODY)-1, -2 and -3). These successes have generated a strong sense of optimism in the genetics community that the same approach holds great promise for identifying genes for a range of common, familial disorders, including those without clear mendelian inheritance patterns. But so far the promise has largely been unfulfilled, as numerous such diseases have proven refractive to positional cloning.

The likely explanation for this is related to the century-old debate between Mendelists and biometricists. The gene mutations studied by Mendel, and those more recently discovered by positional cloning, are those with large effect and strong genotype-phenotype correlations. They are effectively the 'low-hanging fruit' that are easy to harvest. Now, however, we are left with the great majority of the fruit at the top of the tree with no obvious way to reach it. In genetics terms, these are the numerous genes of smaller effect that are likely to underlie most common, familial traits and diseases in humans — that is, the genes more closely related to the biometrical view of the world. Of course, this sharp distinction is artificial, in that in reality gene effects of all magnitudes exist and depend on the

Table 1 Allele sharing Y for ASPs by displacement t , gene frequency p , prevalence K and heritability H

t	p	H_1	$K = 0.001$			$K = 0.01$			$K = 0.10$		
			$H = 0.20$	$H = 0.50$	$H = 0.80$	$H = 0.20$	$H = 0.50$	$H = 0.80$	$H = 0.20$	$H = 0.50$	$H = 0.80$
0.50	0.01	0.001	0.501	0.502	0.502	0.502	0.501	0.501	0.501	0.500	0.500
	0.10	0.011	0.522	0.518	0.514	0.512	0.510	0.508	0.505	0.504	0.503
	0.30	0.026	0.535	0.529	0.524	0.522	0.518	0.515	0.510	0.508	0.507
	0.70	0.026	0.523	0.519	0.516	0.515	0.513	0.511	0.508	0.507	0.506
1.0	0.01	0.005	0.524	0.516	0.512	0.510	0.507	0.506	0.503	0.502	0.502
	0.10	0.043	0.605	0.585	0.569	0.561	0.548	0.539	0.521	0.517	0.515
	0.30	0.095	0.618	0.602	0.589	0.583	0.570	0.560	0.538	0.532	0.528
	0.70	0.095	0.551	0.549	0.544	0.542	0.538	0.534	0.525	0.522	0.520
2.0	0.01	0.019	0.666	0.624	0.594	0.576	0.554	0.541	0.515	0.512	0.509
	0.10	0.153	0.791	0.765	0.741	0.701	0.676	0.656	0.584	0.572	0.563
	0.30	0.296	—	0.734	0.720	—	0.692	0.677	—	0.604	0.595
	0.70	0.296	—	0.585	0.579	—	0.576	0.573	—	0.559	0.555

trait being studied, but it is also true that the larger the gene effect, the less frequent it is likely to be.

The problem can be given a quantitative interpretation by reverting to the model presented above (Fig. 1, Box 1). For complex diseases, linkage analysis is based on the sharing of alleles identical by descent at a marker locus or loci by affected relatives. For pairs of affected sibs, the most frequently used group, it is straightforward to predict the increase in allele sharing for a fully informative marker at or near the disease locus as a function of the genetic model (Box 2).

The observations in Box 2, Table 1 and Fig. 3 provide perspective on results of linkage screens for numerous disorders over the past decade. So far, all genes first identified by linkage analysis and subsequently positionally cloned are those with low allele frequency and high displacement (that is, mendelian or near mendelian inheritance). These include the genes listed above for breast cancer, colon cancer, familial Alzheimer's disease and diabetes. By contrast, no genes with moderate or modest displacement, even for rare disorders, have been identified in this way. The literature is now replete with linkage screens for an array of 'complex' disorders such as schizophrenia, manic-depression, autism, asthma, type 1 and type 2 diabetes, multiple sclerosis and lupus, to name but a few. Although many of these studies have reported significant linkage findings, none has led to convincing replication. Typically, independent studies of the same disorder identify maximal evidence at different chromosomal locations. In effect, linkage analysis, traditionally the most reliable of genetic methods when applied to mendelian traits, has proven to be much less reliable a tool for the study of non-mendelian diseases, with a disappointingly high false-positive rate. The likely explanation is that the biometrical view is closer to reality than the mendelian view for most human traits and diseases.

This does not necessarily mean that no genes underlying non-mendelian traits can be located by linkage analysis. There are several examples of common alleles that have sufficiently large displacement

Table 2 Typology of SNPs and their occurrence

Type	Description	Number (in thousands)
I	Coding, non-synonymous, non-conservative	60–100
II	Coding, non-synonymous, conservative	100–180
III	Coding, synonymous	200–240
IV	Non-coding, 5' UTR	140
V	Non-coding, 3' UTR	300
VI	Other non-coding	> 1,000

to have been detected by linkage analysis. One example is the role of HLA in type 1 diabetes, where allele sharing by affected sib pairs (ASPs) has been estimated at about 73% (ref. 15). A second example is the role of apolipoprotein E (ApoE) in late-onset Alzheimer's disease, where the ASP allele sharing is estimated at about 60%. Other examples probably exist but have yet to be identified, although the number is likely to be few. Table 1 and Fig. 3 indicate that increasing sample sizes may ultimately improve the odds, but there is clearly a limit. In addition, studying more extreme (and less frequent) phenotypes is helpful provided such cases are also genetically more extreme. However, gene effects with displacements of less than 1 standard deviation (s.d.), which are likely to represent most effects, will rarely be identified this way.

These comments apply equally to quantitative traits studied in humans. Designs that select individuals with extreme phenotypes, both concordant for high or low trait values and extremely discordant for high and low trait values, tend to be the most powerful. But again, only loci with high heritabilities or large displacements can be readily identified by linkage analysis^{16,17}.

Another question relates to whether larger families with many affected individuals would provide better power than smaller families, such as sib pairs. The answer depends on the frequency of the susceptibility allele. For high-frequency alleles, selection of dense

Box 2

Linkage evidence as a function of genetic model

Computational details for calculating the expected allele sharing for affected sibs have been given previously¹⁶. I assume the frequency of the high-risk allele is p , the number of standard deviations between homozygotes (displacement) is t , total heritability of liability is H , heritability due to the tested locus is H_1 (and hence residual correlation between sibs is $(H - H_1)/2$) and the population prevalence of disease is K . Table 1 provides the expected allele sharing Y for a pair of affected sibs as a function of the parameters p , t , H and K . (H_1 is derived from p and t as $H_1 = p(1 - p)t^2 / (1 + p(1 - p)t^2)$). As can be seen in Table 1, Y increases directly with t , inversely with K , and inversely with H , but has a more complex relationship with p and H_1 . When the displacement is 0.5, the increase in allele sharing is uniformly minimal. When $t = 1.0$, the increase in allele sharing is again minimal for common traits ($K = 10\%$), moderate for low-frequency traits ($K = 1\%$) when the allele frequency p lies between 10 and 50% and total heritability is low ($< 50\%$), and sizeable for rare traits ($K = 0.1\%$), especially with moderate allele frequencies (10–50%) and low total heritability ($< 50\%$). When displacement is large (2.0), excess allele sharing is high for infrequent and rare traits with moderate allele frequencies (10–50%), but more modest for high allele frequencies ($p = 70\%$). For common traits ($K = 10\%$), the excess allele sharing is sizeable only for moderate allele frequencies. Examination of the locus-specific heritability H_1 shows a complex pattern. Rare alleles with large displacement create low heritabilities but high allele sharing (for example, $p = 0.01$, $t = 2.0$, $H_1 = 1.9\%$), whereas common alleles with large displacement generate high heritabilities but lower allele sharing (for example, $p = 0.70$, $t = 2.0$, $H_1 = 29.6\%$). Thus, genes with low heritability may be identifiable if they are rare and have large displacement but

not if they are common with low displacement. Similarly, genes with high heritability should be detectable unless the allele frequency is high.

Further statistical insight can be obtained by examining expected lod scores as a function of allele sharing Y for different sample sizes. Here I assume N completely informative sib pairs, which gives $2N$ informative identical-by-descent tallies. If the probability of allele sharing is Y , then the probability of R alleles shared out of $2N$ is just $\binom{2N}{R} Y^R (1 - Y)^{2N-R}$. For R shared alleles the maximum lod score is:

$$R \log_{10} R + (2N - R) \log_{10} (2N - R) - 2N \log_{10} N$$

The expected maximum lod score (EMLS) is obtained by taking the sum of lod scores weighted by their probability. The EMLS for sharing values of Y ranging from 0.50 to 0.60 for $N = 250$, 500, 1,000 and 2,000 ASPs is given in Fig. 3. For a genome-wide search, a lod threshold of 3.6 has been recommended⁴⁷. As the figure shows, 250 fully informative sib pairs are sufficient to obtain a significant lod score when $Y \geq 0.60$; 500 sib pairs can detect Y values $\geq 56.5\%$; 1,000 sib pairs can detect values $\geq 55\%$; and 2,000 sib pairs can detect Y values $\geq 53\%$. Comparison with Table 1 shows that genes with displacement of 0.5 or less cannot be detected even with 2,000 sib pairs. Genes with larger displacement (≥ 1.0) should be detectable in many circumstances with 1,000 sib pairs provided the disease is not common ($K \leq 1\%$). Smaller samples (250 sib pairs) have the power to detect genes only with large displacement ($t \geq 2.0$) or genes with intermediate displacement ($t = 1.0$) with intermediate allele frequency (10–50%) for a rare disease ($K = 0.1\%$).

families is likely to increase parental homozygosity at the disease locus and reduce linkage evidence. On the other hand, for rare alleles with large displacement, dense families are usually optimal, because the probability for such a family to be segregating the allele is increased, enhancing the linkage evidence. However, if genome screens of extended pedigrees have been conducted without success, it is reasonable to conclude that rare genes of large effect are unlikely to exist for the trait studied.

Linkage analysis in model systems has actually been far more successful in locating loci with moderate effects for either quantitative traits (quantitative trait loci or QTLs) or disease outcomes than has linkage analysis in humans. There are several reasons for this: (1) inbred strains are often used, which limit the number of loci involved to those that differ between the two strains; (2) rare alleles with large displacement can become fixed in inbred strains subjected to many generations of positive selection; (3) by design, all parents (in an intercross) or half the parents (in a backcross) are heterozygous and thus informative for linkage; and (4) all offspring come from matings of the same phase and thus can be combined into a single large group for analysis. The lack of all of these features in studies of human linkage has probably led to reduced power, but at least some can be addressed by alternate study designs. For example, reducing human genetic variability (item (1) above) is not possible, although focus on certain populations with reduced genetic variation might be beneficial and has been recommended¹⁸. As described above, rare alleles with large displacement in humans can often be identified by studying dense, extended pedigrees (item (2)). Items (3) and (4) above are generally intractable in human linkage studies. The one situation when (3) and (4) apply in humans is when there is linkage disequilibrium (that is, population association) between a marker allele and trait allele. Indeed, when there is complete disequilibrium (or where the trait and marker allele are the same), the human situation becomes directly analogous to the experimental, as individuals from different families can be combined into single groups based on genotype. However, there is still an important difference. In the experimental situation, complete linkage disequilibrium spans the entire length of a chromosome and diminishes only by $(1 - \theta)$ for a marker at recombination fraction θ away from the trait locus in a single experimental generation. In humans, the amount of disequilibrium between a trait allele and marker allele depends on trait allele homogeneity and is a function of the time since the allele first arose and population demographic history over that time. Typically, disequilibrium spans very short chromosome segments except for rare, recent mutations. Finally, it is important to note that, despite the initial success and power of linkage analysis to locate trait loci in model organisms, even in this case positional cloning of these genes has remained a significant challenge.

Back to the future-candidate genes

The disappointing results from linkage studies coupled with a biometrical view of the world has led to the suggestion of alternative approaches to tackling the genetics of non-mendelian diseases, namely reversion to the study of candidate genes on a large scale¹⁹ or high-density genome scans that are dependent on linkage disequilibrium²⁰. However, first it is useful to show directly the greater power of detection of gene effects by direct-association (or linkage-disequilibrium) analysis when the involved variant is in hand as opposed to depending on linkage analysis without linkage disequilibrium (Fig. 4). By using an analysis similar to one described previously¹⁹, ASPs (for linkage) are compared with case-control pairs (for association). Parameterizing the effect of the locus in terms of genotype relative risk (g) and allele frequency (p), for high relative risks ($g \geq 4$) and intermediate allele frequencies ($p = 0.05-0.50$) it is realistic to expect linkage analysis to provide statistical evidence for the location of a disease gene. However, for more modest relative risks ($g \leq 2$), linkage analysis will not provide such evidence except in unrealistically large samples. By contrast, case-control association

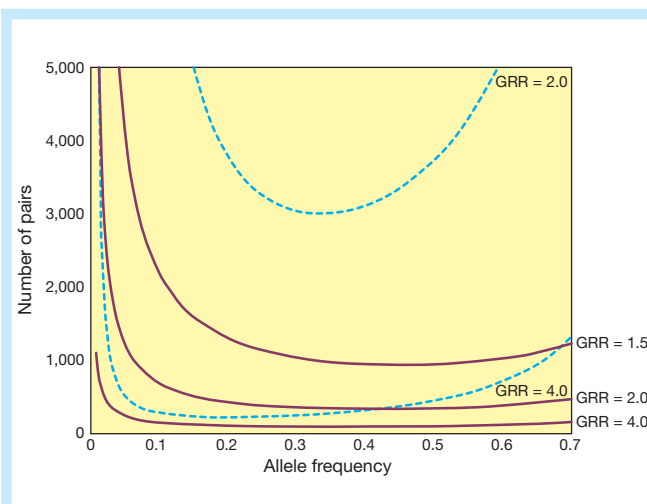


Figure 4 Comparison of linkage (dashed lines) with association analysis (solid lines) for detecting genetic effects. Linkage is based on ASPs with a completely linked and informative marker. Association is based on case-control pairs testing the causative locus. A multiplicative model is assumed, where the genotype relative risk (GRR or g) of the high-risk homozygote is the square of the value of g for the heterozygote, which is given in the figure. Loci with $g > 1.5$ can be detected by association analysis, but $g > 4.0$ is needed to detect a locus by linkage analysis.

studies, even using a stringent significance level (5×10^{-8}), provide adequate power for genes with relative risks as low as 1.5 (with $p = 0.10-0.70$).

Random SNPs or coding SNPs?

The suggestion of genome-wide searches for gene effects using large-scale testing of single nucleotide polymorphisms (SNPs), or perhaps more appropriately simple nucleotide polymorphisms (which could include short deletions and insertions and multinucleotide changes as well as single nucleotide substitutions), has led to considerable discussion of the efficiency of different approaches (see review in this issue by Roses, pages 857–865, for a discussion of SNPs). The original suggestion of Risch and Merikangas¹⁹ was to study coding or promoter variants with potential functional significance. Collins *et al.*²⁰ subsequently suggested that non-coding or evenly spaced SNPs with high density could be used to track disease loci through linkage disequilibrium. The number of SNPs required for the latter strategy has been the subject of debate, primarily because the extent of linkage disequilibrium in the human genome has not been well studied on a large scale. As opposed to recombination — a biological phenomenon already measured extensively in humans — linkage disequilibrium is a property of populations, and thus depends heavily on their demographic and social histories. Population isolates such as Finns, Ashkenazi Jews and Mennonites have been shown to demonstrate extensive linkage disequilibrium (up to several percent recombination) around rare disease mutations. The degree to which the same will be true for higher-frequency variants is uncertain, although as a general rule the disequilibrium is likely to decline with increasing allele frequency owing to an older coalescence time.

Some researchers have argued that as many as 500,000 evenly spaced SNPs may be required to detect linkage disequilibrium of sufficient magnitude for mapping purposes²¹, even in population isolates, whereas others have argued that founder populations, especially those that have remained small over an extended time period, such as the Saami of Scandinavia²² or isolated Sardinian populations²³, would require far fewer SNPs. Although such populations should improve the chances for detecting rare disease alleles (say less than 5% in frequency), owing to greater linkage disequilibrium per base pair, the same is unlikely to be the case for common alleles (greater than 5% in frequency)²⁴. Furthermore, the power of

association tests diminishes significantly with decrease in linkage disequilibrium, and as a result of discordance between the frequencies of disease and marker alleles^{7,25,26}. Although increasing marker density greatly enhances the chance of including a marker in strong linkage disequilibrium with the disease allele, the same is not true for similarity of allele frequencies because correlations between SNP allele frequencies do not increase inversely with distance between SNPs²⁷. Another complication is that, in contrast to linkage analysis, a negative linkage-disequilibrium result in a particular genomic region does not exclude a significant gene effect in that region. It may be that the SNPs used there are in modest or no disequilibrium with the disease allele, and/or the allele frequencies are divergent. Thus, it seems that in a genome-wide random SNP approach, even at high density, many disease-causing genes would be missed.

Several arguments favour using SNPs in coding and promoter regions rather than random SNPs. First, it is these variants, *a priori*, that are most likely to be of functional significance and to influence directly the traits under study. In fact, these are the variants to which random SNP searches are likely to lead. Second, even if not the causative variant in a gene, such SNPs are as likely (or more likely) to be in linkage disequilibrium with the causative allele as are randomly placed SNPs.

Typology of SNPs

If large-scale SNP searches are to become a useful tool for dissecting complex genetic disease, experimental efficiencies need to be brought to bear on the problem. One major efficiency that is possible with association studies but not linkage analysis is DNA pooling, where allele frequencies are examined and compared in a small number of pools rather than a large number of individuals^{7,28–30}. However, it will still be useful to reduce the number of SNPs studied in a systematic way. Although some have argued for an SNP every *n* kilobases (where *n* is between 3 and 100), an alternative approach is to prioritize SNPs based on likely functional significance. The past two decades of study of mendelian traits has provided a rational basis on

which to classify genomic variation (for example, based on the type and frequency of mutations observed for mendelian traits). Two recent studies that have scanned genes for polymorphism^{31,32} also enable estimation of the number of such SNPs in the human genome. The typology and estimated number of SNPs is provided in Table 2. Coding SNPs (or cSNPs) have been denoted as types I to III depending on whether they lead to non-conservative alterations (type I), conservative amino-acid substitutions (type II), or are synonymous (type III). Non-coding SNPs have been separated into 5' untranslated region (UTR) (type IV), 3' UTR (type V) and other non-coding SNPs (type VI). Ultimately, it may be useful to further fragment the last category into subcategories such as exon/intron boundaries and so on.

If we are limited in the number of SNPs to test, it would seem appropriate to give highest priority to type I SNPs (estimated to number between 60,000 and 100,000), as these types of changes are most often associated with functional effects and phenotypic outcomes. In support of this argument, both Cargill *et al.*³¹ and Halushka *et al.*³² found a relative deficiency of SNPs altering amino-acid sequence as compared with synonymous coding or non-coding SNPs, which is consistent with the former having functional and phenotypic significance (and hence subject to selection). Similarly, Halushka *et al.*³² found a relative deficit of allelic diversity in the 5' UTR region of genes, suggesting that type IV SNPs should receive priority (an additional 140,000 SNPs). The same would be true for any variants creating or deleting a splice site.

Another important observation made by Cargill *et al.*³¹ and Halushka *et al.*³² is that type I and II SNPs have lower heterozygosity than other types of SNPs, presumably as a result of selection pressure. For example, Cargill *et al.*³¹ find that about one-quarter of type I and type II SNPs have minor allele frequencies greater than 15%, whereas nearly 60% have minor allele frequencies less than 5%. As discussed below, this observation is important in designing studies to optimize discovery of associations between genes and disease.

Box 3

Glossary

Additive variance. The component of genetic variance due to the additive effects of alleles segregating in the population.

Coalescence time. Number of generations to the most recent common ancestor carrying a mutation or DNA variant currently present in a given population.

Displacement. The difference in s.d. between mean values for individuals with alternative homozygous genotypes at a given locus.

Dominance variance. The component of genetic variance due to non-additive effects of alleles at the same locus.

Epistasis. Genetic variance due to non-additive effects of alleles at distinct loci.

Founder effect. Coalescence of a mutation or DNA variant in a given population to one of the original population founders or his/her descendant.

Genetic heterogeneity. Distinct alleles at the same or different loci that give rise independently to the same genetic disease.

Genotype relative risk (GRR). The risk of disease for one genotype at a locus versus another.

Heritability. The proportion of population variance in a trait attributable to segregation of a gene or genes. Can be

locus-specific or for all loci combined.

Identity by descent. Alleles that trace back to a shared ancestor. For sibs, refers to inheritance of the same allele from a given parent.

Linkage disequilibrium. Two alleles at different loci that occur together within an individual more often than would be predicted by random chance. Also called population allelic association.

Mendelian. A gene with a large displacement, giving rise to a (near) one-to-one correspondence between genotype and phenotype.

Microsatellite. A DNA variant due to tandem repetition of a short DNA sequence (usually two to four nucleotides).

Non-mendelian. A gene without large displacement, giving rise to significant overlap of genotype distributions and lack of one-to-one correspondence between genotype and phenotype.

Restriction-fragment length polymorphism (RFLP). DNA sequence variability leading to cutting or not by a restriction enzyme. Visualized by different patterns of fragment sizes.

Sibling relative risk. The disease risk for a sibling of an affected individual compared to the disease risk in the general population.

Single nucleotide polymorphism (SNP). DNA sequence variation due to change in a single nucleotide.

Table 3 Sample sizes for candidate-gene association studies for different designs

Design		$p = 0.05$		$p = 0.20$	
		No. families	No. subjects	No. families	No. subjects
1	Two controls				
1	Unrelated	872	2,616	300	900
1	Parents	1,251	3,753	417	1,251
1	Sibs (NP)	1,715	5,145	604	1,812
1	Sibs (P)	2,032	6,096	655	1,965
2	Unrelated	265	1,060	102	408
2	Parents	448	1,792	173	692
2	Sibs (NP)	642	2,568	286	1,144
2	Sibs (P)	992	3,968	361	1,444
3	Unrelated	121	605	52	260
3	Parents	218	1,090	101	505
3	Sibs (NP)	314	1,570	177	885
3	Sibs (P)	605	3,025	258	1,290

NP, not pooled; P, pooled.

The typology given above (and in Table 2) is based simply on change in DNA sequence. However, advances in functional genomics/proteomics can also bear on this problem. Discoveries relating to time and distribution of expression of genes, for example deriving from microarray studies, can influence our suspicion of their involvement in various disease processes. It is even conceivable that results of expression studies can be correlated with genotypic variation that exists at a locus. Thus, Table 2 could ultimately be refined to incorporate such information and influence the prioritization of SNPs for phenotype analyses.

Optimal study designs

The recent resurgence of association studies using candidate genes has led to much discussion about design issues. The simplest such design is the epidemiological case-control study, contrasting allele frequencies in cases versus controls. As is true for case-control studies generally, confounding is a problem for inferring a causal relationship between a disease and measured risk factor. One approach to deal with confounding is the matched case-control design, where individual controls are matched to cases on potential confounding factors (for example, age and sex) and the matched pairs are then examined individually for the risk factor to see if it occurs more frequently in the case than in its matched control.

From the genetics perspective, the most serious potential confounder is ethnicity. If cases and controls are not ethnically comparable, then differences in allele frequency will emerge at all loci that differentiate these groups whether the alleles are causally related to disease or not (this phenomenon is sometimes known as stratification artefact). One solution to this problem is to use a matched case-control design, where controls are ethnically matched to cases. This can in theory be accomplished by focusing on homogenous and randomly mating populations, where cases and controls will presumably be ethnically comparable. However, such populations may be more of a theoretical ideal than a reality, as non-random mating patterns exist in nearly all groups. Nonetheless, association studies in Finland are less likely to be subject to confounding problems than in heterogeneous North American populations.

Another solution to this problem involves the use of relatives as controls for cases. The first such described design proposed the use of unaffected sibs as controls^{2,3}, and this design has recently seen a resurgence of interest⁴⁻⁸. Designs involving parents as controls have also been proposed¹³⁻³⁶. Among these, perhaps the test most similar in spirit to the epidemiological matched case-control analysis is the transmission disequilibrium test³⁵, in which an allele transmitted by a parent to an affected child is matched to the other allele not transmitted from the same parent; MacNemar's chi-square test of discordance is then applied to the resulting pairs³⁴ (Fig. 5). The two alleles carried by a parent are of necessity ethnically matched, and thus the stratification artefact is eliminated. The same applies to sib controls, whose genotypes are ethnically matched to the cases.

But a significant result from a design using parent or sib controls still does not imply a causal relationship between the tested allele and the disease outcome, because linkage disequilibrium with a linked locus (but not an unlinked locus) will also create a positive result. Nevertheless, it does at least indicate a significant gene effect nearby, if not the tested allele itself. The main drawback of using parents or sibs as controls is either unavailability (for example, with parents for a late-onset disease) and loss of power, especially with sibs (as described below).

Whereas the simple case-control design is the mainstay of epidemiology, other family-based approaches are available that are more efficient. In particular, sampling multiplex families, where more than a single individual is affected, can be significantly more efficient than sampling singletons. The increase in efficiency is also a function of the disease allele frequency, and is most pronounced for rarer alleles. Using previously described methods^{7,37}, I have calculated the number of families and total individuals required to detect a gene effect with $g = 4.0$ (for the homozygote) and $g = 2.0$ (for the heterozygote), assuming a significance level $\alpha = 5 \times 10^{-8}$ and power $1 - \beta = 80\%$. I evaluate two disease allele frequencies, 5% and 20%, and consider designs including one, two or three affected sibs, where the (two) control individuals are either the parents of the sibship, unaffected sibs, or unrelated.

For all designs except sibs, the efficiency is approximately the same when affected and control samples are pooled. For sibs, greater efficiency is possible with individual genotyping³⁷, so those cases (pooled versus not pooled) are evaluated separately. The results are provided in Table 3. Rarer alleles (0.05 versus 0.20) are always more difficult to detect, but the number of subjects required can be reduced substantially by increasing the number affected in the sibship. Using unaffected sibs as controls leads to two to five times the required sample size as using unrelated subjects, depending on the number of affected sibs. Using parents leads to a 40–80% increase, again depending on number of affected sibs. The main conclusion is that if disease-susceptibility alleles are typically low frequency (say $\leq 20\%$), multiplex sibships are particularly advantageous; they are also advantageous for more frequent alleles, but the relative advantage is less⁷.

An important remaining question is whether to use parents or sibs as controls and suffer the loss in power (especially with sibs), or use unrelated controls and risk loss of robustness. Population stratification has been invoked numerous times as the cause for an observed high false-positive rate in association studies using candidate genes, yet it has rarely been demonstrated as the culprit³⁸. More likely, it is the lack of a stringent significance level used in such studies that is the problem. If one assumes the prior probability for any particular gene variant to be associated with a disease outcome to be low, most reported significant associations will be false positives.

Figure 5 Example of candidate-gene association analysis using different control groups. The case has two *A* alleles. The parental control (alleles not transmitted to the affected child) is two *a* alleles. Analysing the frequency of *A* among transmitted versus non-transmitted alleles by a chi-square test gives rise to the haplotype relative risk test^{32,33}. Pairing each parent's transmitted allele with the non-transmitted allele and comparing the frequency of the two types of discordant pairs (*A* transmitted, *a* non-transmitted, compared with *a* transmitted, *A* non-transmitted) by McNemar's chi-square test gives rise to the transmission disequilibrium test^{33,34}. The sib control alleles are *A* and *a*, and comparison with the affected sib gives rise to sibship-based tests³⁻⁸. The unrelated control (two *a* alleles) gives rise to a traditional matched case-control analysis.



An attractive alternative to using family-based controls is to use random or unlinked genetic markers typed in the same cases and controls to determine the extent of possible confounding by ethnicity³⁹. In fact, the same markers can also be used to assess the significance of any putative association⁴⁰, or even used to adjust any candidate gene analysis for potential confounding by stratified analysis. Given the proposals for large-scale genotyping, it seems most likely that this approach will ultimately be most efficient.

Population variation and replication

As discussed above, rare variants (< 5% frequency) are most likely to be population specific. In some cases, they may be recent in origin and hence specific to a single founder population or less recent and generally found in one major ethnic group (for example, haemochromatosis mutation C282Y found only in Caucasians⁴¹). These are the variants that are most readily detected by a random SNP linkage-disequilibrium approach, but at the same time potentially least replicable by studying distinct populations. In this case it would be worthwhile to examine the same gene in other populations (or even the same population) for other functional variants that are associated with a similar phenotypic endpoint. Discovery of such alleles provides the strongest evidence for a causal link between the gene and the trait, as is the case with family-specific mutations in mendelian diseases.

Common alleles (> 10% frequency) are more likely to be found globally. If so, a causal association between a candidate SNP and trait outcome should be reproducible in many ethnically diverse populations. However, whereas pan-ethnic replicability provides support for a causal relationship, its absence does not necessarily negate it. It is well known that the same mutation can cause a major disease phenotype in one strain of mouse but no phenotype in a genetically distinct strain. Thus, background factors (genetic and otherwise) differentiating populations can modify the expression of a gene and lead to different levels of association. For example, this seems to be the case for ApoE and Alzheimer's disease, where the association exists pan-ethnically but is strongest in Caucasians and Asians, and weaker in Hispanics and African Americans⁴².

Another advantage to having an ethnically diverse sample of individuals/families is that patterns of linkage disequilibrium may differ ethnically, helping to resolve causal from non-causal relationships. While populations with high linkage disequilibrium may be useful for initial detection of SNP associations, several different SNPs may be in strong or complete disequilibrium. Populations with lower levels of disequilibrium can help resolve which SNP effect is primary. Generally, Africans appear to have the lowest levels of linkage disequilibrium and hence are likely to be most useful for such analyses. An example is provided by the association of HLA and narcolepsy. In Caucasian and Asian populations, the alleles *DR2* and *DQB-0602* are equally associated with the disease (and in complete disequilibrium with each other), whereas in

Africans there is incomplete disequilibrium between them and *DQB-0602* shows the primary effect⁴³.

Conclusions

As we move into a new millennium, the association of computational and molecular technological developments, including the sequencing of the human genome, is opening up new and unprecedented opportunities for genetics research. It is appropriate to reflect on the accomplishments of the past century and where the new technology is likely to lead us.

As I have indicated, much of the current debate in human genetics regarding approaches to the study of complex diseases can be reflected back onto the century-long debate between the Mendelist view and the biometricist view of the world. Much of the difference in views can be attributed to the traits chosen for study, with Mendelists focusing on those dominated by single-gene effects and the biometricists focusing on continuous, 'polygenic' variation. For most common diseases facing humanity, it is likely that the biometrical view is more apt.

The past two decades have witnessed numerous spectacular applications of positional cloning to identify mendelian human disease genes. But the fact is that the same approach is proving limited in identifying the multitude of genes underlying the more common, complex disorders. Even high-density genome scans with evenly spaced SNPs, depending on linkage disequilibrium, are simply an extension of the same reverse-genetics approach.

At this turn of the millennium, with the completion of the human genome project now in sight, we need to consider the full impact of having the entire human DNA sequence. Although the traditional reverse-genetics approaches (linkage and linkage-disequilibrium analysis) may identify a few of the genetic susceptibility agents we seek, I believe a far greater yield will occur by rethinking this problem from a forward-genetics perspective. Identifying all (or most) of the genes in the human genome, as well as identifying and cataloguing the functional variation lying within them, which occurs naturally in the human population, provides opportunities for studying the impact of those variants on phenotypic outcomes of interest. Functional genomics technology involving microarrays and proteomics will provide added insights regarding gene function on the cellular level, improving our ability to predict phenotypic effects of genes at the organismic level. Nevertheless, efficient study designs will still be required, and multiplex families, the mainstay of linkage-based studies, will still be optimal. However, instead of family-based controls, unrelated controls will emerge as a more powerful and efficient approach (especially for analyses based on pooled DNA samples), and robustness will be maintained by studying a large number of independent SNPs. Sampling families of varying ethnicity will also be advantageous from the perspective of enhancing evidence of causality as well as identifying genetic and/or environmental modifying factors.

Despite future developments, it will still be important to view the study of human disease from an epidemiological perspective. Both human genetics and epidemiology are observational as opposed to experimental sciences, and we will never be able to exert the degree of scientific control in studies of human disease that experimentalists can with model systems. Furthermore, we must not lose sight of the numerous non-genetic influences that influence disease risk, and how they interact with host (that is, genetic) factors. □

1. Vogel, F. & Motulsky, A. G. *Human Genetics: Problems and Approaches* (Springer, Berlin, 1982).
2. Penrose, L. S. Some practical considerations in testing for genetic linkage in sib data. *Ohio J. Sci.* **39**, 291–296 (1939).
3. Clarke, C. A. *et al.* ABO blood groups and secretor character in duodenal ulcer. *Br. Med. J.* **2**, 725–731 (1956).
4. Curtis, D. Use of siblings as controls in case-control association studies. *Am. J. Hum. Genet.* **61**, 319–333 (1997).
5. Spielman, R. S. & Ewens, W. J. A sibship based test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.* **62**, 450–458 (1998).
6. Boehnke, M. & Langefeld, C. D. Genetic association mapping based on discordant sib pairs: the discordant-alleles test. *Am. J. Hum. Genet.* **62**, 950–961 (1998).
7. Risch, N. & Teng, J. The relative power of family-based and case-control designs for association studies of complex human diseases. I. DNA pooling. *Genome Res.* **8**, 1273–1288 (1998).
8. Schaid, D. J. & Rowland, C. Use of parents, sibs and unrelated controls for detection of associations between genetic markers and disease. *Am. J. Hum. Genet.* **63**, 1492–1506 (1998).
9. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
10. Weber, J. L. & May, P. E. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**, 388–396 (1989).
11. Litt, M. & Luty, J. A. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* **44**, 397–401 (1989).
12. Rao, D. C., Keats, B. J. B., Morton, N. E., Yee, S. & Lew, R. Variability of human linkage data. *Am. J. Hum. Genet.* **30**, 516–529 (1978).
13. Morton, N. E. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.* **7**, 277–318 (1955).
14. Ott, J. *Analysis of Human Genetic Linkage* (Johns Hopkins University Press, Baltimore, 1991).
15. Concannon, P. *et al.* A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nature Genet.* **19**, 292 (1998).
16. Risch, N. & Zhang, H. Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**, 1584–1589 (1998).
17. Eaves, L. & Meyer, J. Locating human quantitative trait loci: guidelines for the selection of sibling pairs for genotyping. *Behav. Genet.* **24**, 443–455 (1994).
18. Terwilliger, J. D., Zollner, S., Laan, M. & Paabo, S. Mapping genes through the use of linkage disequilibrium generated by genetic drift: “drift mapping” in small populations with no demographic expansion. *Hum. Hered.* **48**, 138–154 (1998).
19. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
20. Collins, F. S., Guyer, M. S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
21. Kruglak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1998).
22. Laan, M. & Paabo, S. Demographic history and linkage disequilibrium in human populations. *Nature Genet.* **17**, 435–438 (1997).
23. Lonjou, C., Collins, A. & Morton, N. E. Allelic association between marker loci. *Proc. Natl Acad. Sci. USA* **96**, 1621–1626 (1999).
24. Wright, A. F., Carothers, A. D. & Pirastu, M. Population choice in mapping genes for complex diseases. *Nature Genet.* **23**, 397–404 (1999).
25. Muller-Myhsok, B. & Abel, L. Genetic analysis of complex diseases. *Science* **275**, 1328–1329 (1997).
26. Tu, L.-P. & Whittemore, A. S. Power of association and linkage tests when the disease alleles are unobserved. *Am. J. Hum. Genet.* **64**, 641–649 (1999).
27. Nickerson, D. A. *et al.* DNA sequence diversity in a 9.7kb region of the human lipoprotein lipase gene. *Nature Genet.* **19**, 233–240 (1998).
28. Arnheim, N., Strange, C. & Erlich, H. Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci. *Proc. Natl Acad. Sci. USA* **82**, 6970–6974 (1985).
29. Carmi, R. *et al.* Use of DNA pooling strategy to identify a human obesity syndrome locus on chromosome 15. *Hum. Mol. Genet.* **3**, 1331–1335 (1995).
30. Barcellos, L. F. *et al.* Association mapping of disease loci by use of a pooled DNA genomic screen. *Am. J. Hum. Genet.* **61**, 734–747 (1997).
31. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
32. Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
33. Falk, C. T. & Rubinstein, P. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann. Hum. Genet.* **51**, 227–233 (1987).
34. Terwilliger, J. D. & Ott, J. A haplotype-based “haplotype-relative risk” approach to detecting allelic associations. *Hum. Hered.* **42**, 337–346 (1992).
35. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516 (1993).
36. Thomson, G. Mapping disease genes: family-based association studies. *Am. J. Hum. Genet.* **57**, 487–498 (1995).
37. Teng, J. & Risch, N. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res.* **9**, 234–241 (1999).
38. Morton, N. E. & Collins, A. Tests and estimates of allelic association in complex inheritance. *Proc. Natl Acad. Sci. USA* **95**, 11389–11393 (1998).
39. Pritchard, J. K. & Rosenberg, N. A. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65**, 220–228 (1999).
40. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
41. Feder, J. N. *et al.* A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature Genet.* **13**, 399–408 (1996).
42. Farrer, L. A. *et al.* Effects of age, sex and ethnicity on the association between apolipoprotein E genotype and Alzheimer’s disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *J. Am. Med. Assoc.* **278**, 1349–1356 (1997).
43. Mignot, E. *et al.* DZB1*0602 and DQA1*0102 (DQ1) are better markers than DR2 for narcolepsy in Caucasian and Black Americans. *Sleep* **17**, S60–S67 (1994).
44. Kempthorne, O. *An Introduction to Genetic Statistics* (Iowa Univ. Press, Ames, 1969).
45. Khoury, M. J., Beaty, T. H. & Cohen, B. H. *Fundamentals of Genetic Epidemiology* (Oxford University Press, New York, 1993).
46. Risch, N. Linkage strategies for genetically complex traits. I. Multi-locus models. *Am. J. Hum. Genet.* **46**, 222–228 (1990).
47. Lander, E. & Kruglak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* **11**, 241–247 (1995).

Pharmacogenetics and the practice of medicine

Allen D. Roses

Genetics Directorate, Glaxo Wellcome plc, Greenford, Middlesex UB6 0HE, UK, and Duke University Medical Center, Durham, North Carolina 27710, USA

"If it were not for the great variability among individuals medicine might as well be a science and not an art." The thoughts of Sir William Osler in 1892 reflect the view of medicine over the past 100 years. The role of physicians in making the necessary judgements about the medicines that they prescribe is often referred to as an art, reflecting the lack of objective data available to make decisions that are tailored to individual patients. Just over a hundred years later we are on the verge of being able to identify inherited differences between individuals which can predict each patient's response to a medicine. This ability will have far-reaching benefits in the discovery, development and delivery of medicines. Sir William Osler, if he were alive today, would be re-considering his view of medicine as an art not a science.

Every individual is a product of the interaction of their genes and the environment. Pharmacogenetics is the study of how genetic differences influence the variability in patients' responses to drugs. Through the use of pharmacogenetics, we will soon be able to profile variations between individuals' DNA to predict responses to a particular medicine. The medical significance and economic value of a simple, predictive medicine response profile, which will provide information on the likelihood of efficacy and safety of a drug for an individual patient, will change the practice and economics of medicine. The ability to rapidly profile patients who are likely to benefit from a particular medicine will also streamline drug development and provide opportunities to develop discrete medicines concurrently for different patients with similar disease phenotypes. Other than relatively rare and highly penetrant diseases related to mutations of a single gene inherited in families (Box 1), science has never before had the tools to characterize the nuances of inherited metabolic variations that interact over time and lead to common diseases. Powerful pharmacogenetic research tools are now

becoming available to classify the heterogeneity of disease as well as individual responses to medicines.

An ongoing ethical debate concerning potential genetic applications and the impact on individuals and families accompanies scientific advances. Clearly defined terminology should form the basis for informative discussions so that the word 'genetics' is not demonized. For example, tests that are specific to disease genes can help diagnose disease, determine the carrier status of an individual or predict the occurrence of disease. These are quite distinct from profiles that, for example, are specific for genes involved in drug metabolism, which provide information on how a medicine will be metabolized in an individual. In the near future (1–3 years) there will be non-disease- and non-gene-specific pharmacogenetic profiles developed to determine whether an individual is likely to respond to a medicine and/or to not experience serious side effects. Language needs to be more precise so that there can be clarity, especially for public policy debates. Pharmacogenetics is not gene therapy, not genetically modified foods, not genetic engineering, and not cloning of humans or their organs. Ethical, legal and social implications for 'genetic tests' of single-gene mutational diseases should

Box 1

Genes and disease

Single-gene mutations and rare diseases (mendelian inheritance)

Causally related to rare inherited diseases (high penetrance)
Examples:

- Cystic fibrosis (*CFTR* gene)
 - Inheritance: autosomal recessive
 - Location: chromosome 7 (7q31)
 - Mutation: deletion of 3 bp at codon 508 accounts for 70% of mutations
- Huntington disease:
 - Inheritance: autosomal dominant
 - Location: chromosome 4 (4p16.3)
 - Mutation: cytosine/adenine/guanine repeat >35 times

Gene polymorphisms and common disease susceptibility (polygenic; complex inheritance)

Examples of susceptibility loci:

- Common late-onset Alzheimer's disease
 - Susceptibility gene (*ApoE*) on chromosome 19 (19q13)
 - Susceptibility gene locus on chromosome 12 (12q)
- Migraine
 - Susceptibility gene loci on chromosome 19 (19p13); chromosome X (Xq24)
- Non-insulin-dependent diabetes mellitus
 - Susceptibility gene loci on chromosome 12 (12q); chromosome 2 (2q)
- Psoriasis
 - Susceptibility gene locus on chromosome 3 (3q21)

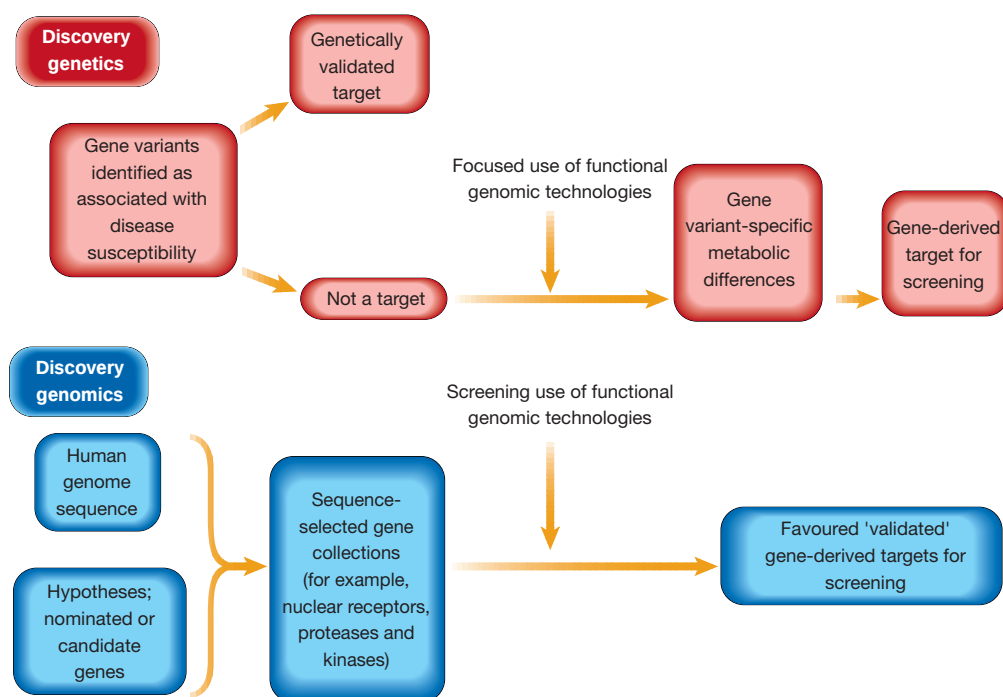


Figure 1 Genetics and genomics to identify drug targets. Two general strategies are used to identify genes and find new targets for drugs: genetics and genomics. Each approach shares technologies, like functional genomics, but as a part of different experimental designs. Genetics identifies disease-related susceptibility genes and genomics identifies genes that belong to similar families based on their sequence homologies. The goal of most genomic strategies is to collect genes that may be expressed and used for high-throughput screening targets. Any one of the identified genes may or may not have a connection to any disease process, with a high probability

that it does not. Focused uses of functional genomic technologies include, for example, study of lines of transgenic mice that differ only in the specific polymorphisms defined in the susceptibility gene that relates to disease expression in humans. Understanding isoform-specific metabolic functions can lead to the identification of new metabolic targets for drug screening. Screening use of functional genomic technologies are used to imply validation for targets derived from discovery genomics, such as higher gene expression in a tissue of a subset of genes or the expression of protein observed in disease tissues but not seen in comparable tissue from controls.

not automatically be assumed for other non-disease-specific applications simply because they are labelled imprecisely as 'genetic tests'. Use of inaccurate terminology may hinder and delay the significant health-care benefits that will accrue from pharmacogenetics.

It is important to discuss how the benefits of pharmacogenetics can be applied to drug development and the provision of better health care today — 3–5 years before the widespread application of pharmacogenetics. This will enable the maximum benefits for patients to be obtained as rapidly as possible. In this review I begin with a brief discussion of how genetics and genomics are used in the pharmaceutical industry to identify targets and discover new medicines that will stop or prevent disease processes and then discuss how pharmacogenetics will impact the pharmaceutical industry and the provision of health care.

Target selection

Target validation that will predict a well-tolerated and effective medicine for a clinical indication in humans is a widely perceived problem; but the real challenge is target selection^{1–3}. A limited number of molecular target families have been identified, including receptors and enzymes, for which high-throughput screening is currently possible. A good target is one against which many compounds can be screened rapidly to identify active molecules (hits). These hits can be developed into optimized molecules (leads), which have the properties of well-tolerated and effective medicines. Selection of targets that can be validated for a disease or clinical symptom is a major problem faced by the pharmaceutical industry. The best-validated targets are those that have already produced well-tolerated and effective medicines in

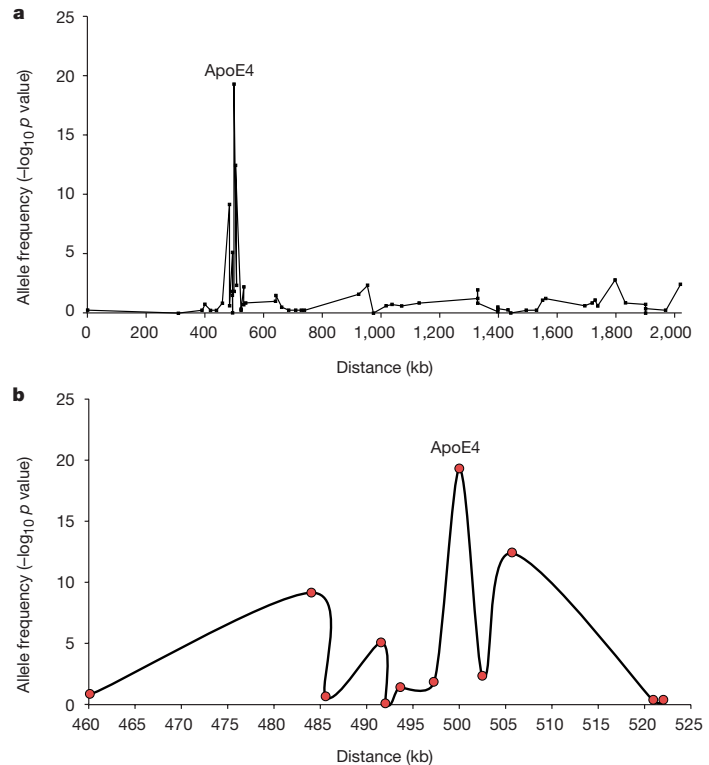
humans (precedented targets). Many targets are chosen on the basis of scientific hypotheses and do not lead to effective medicines because the initial hypotheses are often subsequently disproved.

Two broad strategies are being used to identify genes and express their protein products for use as high-throughput targets. These approaches of genomics and genetics share technologies but represent distinct scientific tactics and investments. Discovery genetics uses human disease populations to identify disease-related susceptibility genes. Discovery genomics uses the increasing number of databases of DNA sequence information to identify genes and families of genes for tractable or screenable targets that are not known to be genetically related to disease.

The advantage of information on disease-susceptibility genes derived from patients is that, by definition, these genes are relevant to the patients' genetic contributions to the disease. However, most susceptibility genes will not be tractable targets or amenable to high-throughput screening methods to identify active compounds^{1,3}. The differential metabolism related to the relevant gene variants can be studied using focused functional genomic and proteomic technologies to discover mechanisms of disease development or progression. Critical enzymes or receptors associated with the altered metabolism can then be used as targets. Gene-to-function-to-target strategies that focus on the role of the specific susceptibility gene variants on appropriate cellular metabolism become important (Fig. 1).

Data mining of sequences from the Human Genome Project and similar programmes with powerful bioinformatic tools has made it possible to identify gene families by locating domains that possess similar sequences. Genes identified by these genomic strategies

Figure 2 Significance of SNP allele frequency differences in an affected Alzheimer's disease population and age-matched controls. **a**, The association data for dozens of ordered SNPs from a region of 2 million bases on either side of *ApoE*. When the allele frequencies of each SNP are compared in large series of Alzheimer's disease patients and controls, a sharp peak of several SNPs can be readily observed in linkage disequilibrium, with no significant difference in the frequencies of background alleles. (From ref. 45; published with permission.) **b**, If the peak is enlarged to illustrate a region of only ~60,000 bases around *ApoE*, three SNPs from the map that are each highly significantly associated with Alzheimer's disease can be identified. Only two genes, *ApoC1* and *ApoE*, are coded in the physical DNA segment defined by the SNPs associated with Alzheimer's disease. The association data from the SNP defining the *ApoE4* polymorphism, known to be associated with earlier onset of disease, is also illustrated. Not illustrated is the lack of association of another defined *ApoE* polymorphism, that for *ApoE2*. *ApoE2* is associated with protection, or later onset of the phenotype of Alzheimer's disease. Although *ApoE2* is in linkage disequilibrium with *ApoE4*, there is no association with the disease^{7,59}. Thus whereas SNPs may be in linkage disequilibrium, as are those for *ApoE2* and *ApoE4*, the association with Alzheimer's disease is found only for several SNPs in linkage disequilibrium with *ApoE4* and Alzheimer's disease. It is the presence of these SNPs that allow rapid recognition of the region within which *ApoE4* is located. (From ref. 45; published with permission.)



generally require some sort of functional validation or relationship to a disease process. Technologies such as differential gene expression, transgenic animal models, proteomics, *in situ* hybridization and immunohistochemistry are used to imply relationships between a gene and a disease process. Over the next five years there will be many opportunities to identify the full complement of gene families. Some of these families can provide tractable targets for high-throughput screening of molecules.

The difference between the genomic approach and the genetic approach is that the former creates a need to functionally validate the tissue distribution and other aspects of each identified gene and find a relevant disease or clinical indication. In contrast, once the disease-related variants of susceptibility disease genes are identified, a single susceptibility gene is automatically validated in human disease. The major distinction between the genomic and genetic approaches is target selection, with genetically defined genes and variant-specific targets already known to be involved in the disease process. The current vogue of discovery genomics for nonspecific, wholesale gene identification, with each gene in search of a relationship to a disease, creates great opportunities for development of medicines. However, there are also enormous economical costs associated with searching huge lists of genes for 'the right disease for the available gene'. It is correct to state that target validation is a major challenge to the pharmaceutical industry, but it is also critical to realize that the core problem for drug development is poor target selection. The screening use of unproven technologies to imply disease-related validation, and the huge investment necessary to progress each selected gene to proof-of-concept in humans, is based on an unproven and cavalier use of the word 'validation'. Each failure is very expensive in lost time and money.

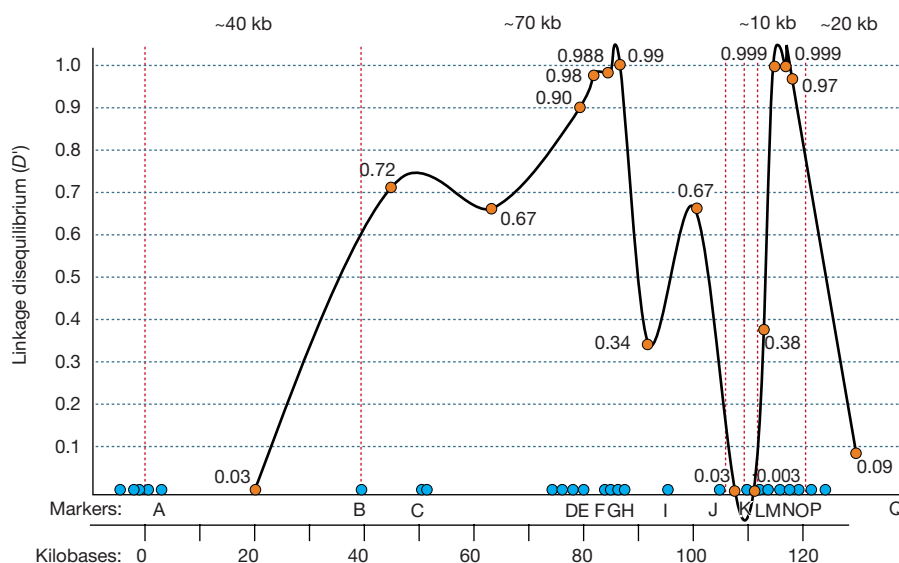
For example, differential gene expression (DGE) and proteomics are screening technologies that are widely used for target validation. They detect different levels and/or patterns of gene and protein expression in tissues, which may be used to imply a relationship to a disease affecting that tissue⁴⁻⁶. Screening with these powerful tools has yet to lead to a specific target for a drug candidate with proven efficacy in humans or to a marketed drug. In fact, the proof-of-concept experiments to demonstrate that differences in the tissue expression

of a particular gene are related to disease expression (two very different meanings to 'expression') have not been performed in any common disease with known susceptibility genes. Neither have functional genomic screening methods yet been applied to rare mutational diseases for proof of principle. Rather there has been a tacit and widespread assumption that differentially expressed genes will be related causally to disease progression, rather than as a consequence of disease-related processes. Selecting the right gene using large-scale screening technologies is a significant and expensive problem.

There at least are two common disease examples in which the expression of genetic differences identified by DGE technologies would not have led to target definition. The gene encoding apolipoprotein E (*ApoE*) is a known susceptibility gene for common, late-onset Alzheimer's disease. Specific allelic variants that are inherited determine the risk and age of onset distribution of the disease^{7,8}. Traditional tissue immunohistochemical and *in situ* hybridization studies of the distribution of *ApoE* have been more revealing than functional genomic screening methods, showing that *ApoE* is expressed in human neurons under normal conditions, but not in rodent neurons, which are used to model characteristics of Alzheimer's disease⁹⁻¹². Differential expression of total brain *ApoE* in patients with Alzheimer's disease has not led to the identification of tractable targets. It is highly unlikely (and to date untested as a proof of principle) that β -amyloid precursor protein (APP) or presenilin mutations, each causing rare, early-onset, dominantly inherited Alzheimer's disease, would have been identified using these methods. Yet DGE and proteomic screening methods are currently major investments in several research programmes that work on Alzheimer's disease.

The converse experiment was published recently for an already validated target. Peroxisome proliferator-activated receptor- γ (PPAR- γ) is a nuclear receptor with documented involvement in glucose metabolism and homeostasis^{13,14}. PPAR- γ can be considered a precedent target molecule, which can be screened using high-throughput methods for molecules that are effective in treating diabetes mellitus. In this case, there was no previous evidence for PPAR- γ as a susceptibility gene for diabetes mellitus, nor was there any

Figure 3 Linkage disequilibrium data for 12 adjacent SNPs that are located and ordered within the 120-kilobase region encoding a migraine susceptibility gene (in this instance, a D' value above 0.30 is indicative of highly significant linkage disequilibrium). Five of these 12 SNPs also demonstrated significant association with migraine, illustrating the use of linkage disequilibrium mapping to identify disease-associated polymorphisms.



abnormality in differential PPAR- γ expression. But a rare and severe form of diabetes mellitus has now been shown to be related to specific mutations of the PPAR- γ molecule¹⁵, thus providing further validation of PPAR- γ as a target. There is, however, no indication that DGE screening or proteomic analyses of comparative tissues from common diabetic patients would have identified the preceded molecular target, PPAR- γ . In this case, the genetic data followed validation of a target in humans, and not from differential genomic screening techniques.

The identification of disease-susceptibility genes and study of the function of the susceptibility gene variants will lead to targets that, by definition, will be related to the disease in patients and will therefore be validated. This process identifies few targets compared with the approach used in discovery genomics of data-mining human sequence information. It is therefore practical to use both genetic and genomic strategies and to focus screening technologies to 'pick the winners'.

Pharmacogenetics and medical practice

Diagnosis

When we go to see our doctor, our symptoms and physical signs are evaluated, and appropriate tests (for example, blood, urine, X-ray and magnetic resonance imaging) are undertaken. To the non-physician, this process of disease diagnosis seems straightforward. However, for a patient to have all the classical symptoms and signs of a particular disease is the exception rather than the rule. How these diagnoses relate to the underlying mechanism of disease is often unknown. For example, patients with mutations in different genes may present as clinically identical. Mutations of APP, presenilin 1 and presenilin 2 lead to clinically indistinguishable forms of Alzheimer's disease^{16–19}. It is also important to note that mutations at different sites along the APP gene can lead to two distinct diseases, early-onset Alzheimer's disease and recurrent intracerebral haemorrhages²⁰. For many common diseases, the situation may be assumed to be even more complicated, with many contributing molecular variants of several interacting susceptibility genes leading to multiple clinical effects over varying time frames^{7,21–25} (Box 1). Thus many of the diseases that we classify clinically may be syndromes with several distinct contributing pathogenic mechanisms. With all this clinical and genetic heterogeneity we should not lose sight of the fact that the major objective is to treat, cure or prevent disease. It is significant that a medicine works; does it matter whether it is effective in patients who may have different diagnoses? The goal of medicine is to relieve pain and suffering. Similar mechanisms may exist for quite diverse clinical diseases. As the targets and mechanisms

are validated in humans, additional clinical indications may become more obvious because of shared mechanisms rather than similar clinical presentations.

How does your doctor know when making the diagnosis that medicines that are effective for you have not been precluded? Pharmacogenetics will enable individuals to be classified according to their likely response to a medicine. This is not a new concept as clinical subtypes are often classified by drug responsiveness (for example, steroid-sensitive and steroid-resistant asthma). Application of pharmacogenetics will expand the population to those who can be helped but might have otherwise been missed because their clinical syndrome did not fit neatly into a traditional disease category. Alosetron is a recently approved medicine in the United States for the treatment of female patients with diarrhoea-predominant irritable bowel syndrome (IBS)^{26,27}. Most physicians will acknowledge that the diagnosis of IBS can be imprecise — in fact, the 'disease' is truly a syndrome. The value of a diagnostic test to sub-classify IBS into different types may be limited, but a simple medicine response profile to determine whether the patient's symptoms will be alleviated by alosetron could have considerable value²⁸. Pharmacogenetic approaches will no doubt confirm what clinicians already know — disease diagnosis is not easy nor necessarily homogeneous and accurate.

Apparently distinct diseases may have similar underlying mechanisms. A medicine developed for a specific indication could have value in treating other related or non-related conditions. This is also not a new concept. There are many medicines that were initially registered with a single indication, which have then been expanded as more clinical research is conducted. For example, carbamazepine was initially registered as a treatment for trigeminal neuralgia, a syndrome with intermittent severe lightning-like bursts of facial pain, but was later extended to treat various forms of epilepsy. By understanding the genetic basis of patient responses to medicines, and perhaps also by having a better understanding of how the medicine works, we will be able to identify additional clinical indications more quickly.

Treatment

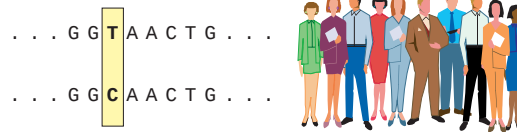
How does a physician know if the medicine and the dose prescribed will be effective and whether or not the patient will experience adverse effects? Information is available from clinical trials in the medicine's data sheet/label in which similar patients were included and the physician may use experience of treating previous patients. On many occasions, the prescribed medicine will be effective and not cause serious side effects. Other patients may not respond or suffer adverse reactions. By applying the results of pharmacogenetic

Box 2

SNPs and pharmacogenetics

What is an SNP?

Different people can have a different nucleotide or base at a given location on a chromosome

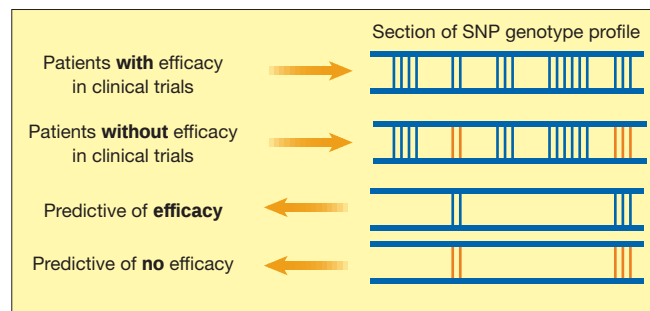


What is an SNP map?

Location of SNPs on human DNA

Human DNA

How can an SNP map be used to predict medicine response?



research to clinical practice, physicians will be able to use information from patients' DNA to determine how patients are likely to respond to a particular medicine. The clinical fact that the drug dose for some patients must be individualized has been accepted for years. Polymorphisms in genes encoding P450 enzymes, *N*-acetyltransferase and other key enzymes in drug metabolism account for the concentration variation of certain drugs in patients' blood^{29,30}. It is also well established that some patients can be slow in activating drugs and respond inadequately to some prodrugs, or exhibit reduced clearance and increased effects from some pharmacologically active agents^{31–33}. Enzyme tests that measure those variants have, in some cases, already been replaced with genetic variants on chips. In the future, metabolic screens of genetic variants will be standardized so that automated read-outs of each person's predicted response to each medicine could be generated. These DNA-based screens will not provide disease-specific diagnosis, but useful information to aid in individual dosing of medications or avoidance of side effects.

SNP mapping: a tool for personalized genetic profiling

Single nucleotide polymorphisms (SNPs) are single-base differences in the DNA sequence that can be observed between individuals in the population^{34–36} (Box 2). A polymorphism has been defined as the least common allele occurring in 1% or greater of the population³⁷, whereas mutations are rare differences which occur in less than 1% of the population (usually much less than 1%). Typically, mutations have been discovered in coding sequences of genes causing rare inherited diseases³⁸. SNPs are present throughout the human genome with an average frequency of approximately 1 per 1,000 base pairs (bp)³⁵. 'The SNP Consortium' (a consortium of pharmaceutical and bio-informational companies, five academic centres and a charitable trust) is currently producing an ordered high-density SNP map of the human genome (Box 2). Mapped SNPs are being placed regularly into public domain websites (<http://snp.cshl.org>). The original target was to produce an SNP map with 200,000–300,000 SNPs evenly distributed throughout the human genome. In fact, this initiative is ahead of schedule and will probably provide 600,000–800,000 SNPs by the end of year 2 (April 2001). This map

will enable disease and drug response phenotypes to be mapped by linkage disequilibrium. Linkage disequilibrium occurs when haplotype combinations of alleles at different loci occur more frequently than would be expected from random association; it decays with time (generations) in proportion to the recombination fraction between the loci. When alleles are physically close, they are more likely to be inherited together than are alleles that are further apart. Therefore, variations of several ordered SNP markers that are close to, or within, a particular gene variant on a chromosome are likely to be inherited together with that gene variant when they are in linkage disequilibrium. So consecutive SNP variations that are in linkage disequilibrium and associated with a disease phenotype can 'mark' the position on the chromosome where a susceptibility gene is located.

Recent data show the utility of using high-density SNP linkage disequilibrium mapping to find disease-susceptibility genes. Before these SNP mapping experiments, individual testing of multiple candidate genes found to be located within a linkage region was a long, expensive and relatively unproductive way of searching for disease-susceptibility genes.

Polymorphisms of the *ApoE* gene provided the first proof of principle for the detection of a linkage disequilibrium locus around a known susceptibility gene for Alzheimer's disease. In 1997, a high-density SNP map for a region of 4 million bases (0.1% of the human genome) around the *ApoE* locus on chromosome 19 was constructed³⁹. The goal of the experiment was to determine whether the *ApoE* gene could be detected as a susceptibility locus associated with Alzheimer's disease using high-density SNP mapping to detect a small region of linkage disequilibrium (Fig. 2a,b)⁴⁰. These studies showed that by using DNA from patients with Alzheimer's disease and controls, it is possible to detect those SNPs in linkage disequilibrium that are associated with the disease.

This methodology has been used to identify susceptibility genes for other diseases such as migraine with aura, which is also localized on chromosome 19 (Fig. 3). In this case a linkage region of approximately 1 million bases was reduced to a 70,000–120,000-bp locus (C.-F. Xu *et al.*, unpublished results). Although this linkage disequilibrium segment of DNA is larger than that found for *ApoE* and

Alzheimer's disease, it contains the coding sequences of a single gene. Similar data have been collected and tested for psoriasis on chromosome 3 (C.-F. Xu *et al.*, unpublished results) and non-insulin-dependent diabetes mellitus on chromosome 12 (E. Lai *et al.*, unpublished results). Thus it is now possible to rapidly reduce the size of the DNA region which contains disease-susceptibility genes by two to three orders of magnitude from millions of base pairs to thousands of base pairs. In practical terms, this accelerates the identification of susceptibility genes within the relatively large regions of DNA that are found by traditional linkage using typical 400 marker screens. Using theoretical, simulated data, some researchers had suggested that one SNP per 6,000 bp would be necessary to locate disease-susceptibility genes⁴¹. These simulations have been questioned and are not supported by published data or data from mapping of susceptibility genes^{42,43}. In particular, this research in these disease areas is a practical demonstration that a density of SNPs of one every 10,000–30,000 bp can rapidly narrow the search for susceptibility genes. After the high-density SNP map of the whole genome is completed, it will no longer be necessary to create SNP maps at each disease locus as they will already exist. Thus the rate of discovery of susceptibility genes will depend on the quality of the patient and control populations, rather than being limited by the technical capacity to construct new, ordered, limited SNP maps. The next technical hurdle will be the development of inexpensive high-throughput methods for scoring large numbers of SNPs from hundreds of patients and controls. Considerable efforts are now underway within the biotechnical community to establish low-cost, high-throughput, accurate SNP scoring technologies.

Determining abbreviated SNP linkage disequilibrium profiles

SNPs are the simplest form of DNA polymorphism. Using currently available DNA analysis systems, such as chip-based resequencing or microsphere-based analytical methodologies, thousands of SNPs can be read out automatically and rapidly^{36,44}. By applying whole-genome SNP linkage disequilibrium mapping to patients during phase II clinical trials of a medicine, it may be possible to select multiple small regions from the whole-genome SNP map where SNPs are in linkage disequilibrium and associated with efficacy and common adverse event phenotypes⁴⁵. Selecting only these small regions of SNP linkage disequilibrium into abbreviated SNP linkage disequilibrium profiles (Box 2) will enable more rapid and inexpensive screening of patients who are likely to experience efficacy or adverse events in response to that medicine⁴⁶. Thus whereas the phase II SNP scan might genotype 200,000 SNPs for each patient, the critical data used for identifying markers for efficacy for subsequent phase III clinical trials may use only several hundred SNPs from multiple small regions in linkage disequilibrium and associated with efficacy or adverse events. The abbreviated patterns for efficacy could be extended during large-scale post-approval drug surveillance (see below) to include further efficacy phenotypes and adverse event profiles without providing any significant collateral disease information for relatives regarding inheritance of any specific disease-associated gene allele.

Chip technologies are already available for accurately genotyping hundreds to a few thousand SNPs³⁶. The cost of chips as a platform for medicine response profiling is likely to be reduced when analyses of hundreds of thousands of patients are performed once the medicine is marketed. In fact, each chip could contain a panel of abbreviated SNP linkage disequilibrium profiles for several drugs with the same clinical indications so that the most appropriate medicine with that indication for that patient can be determined from a single blood sample.

Similar analyses of patients with identical disease phenotypes could be used to determine disease heterogeneity. Different SNP linkage disequilibrium profiles of patients with the same disease phenotype could define patterns of disease heterogeneity without necessarily identifying the actual genes and alleles involved^{36,47}. Genetic research

conducted during phase II clinical trials of investigational medicines could use the high-density SNP map of the human genome to identify the sub-type of the disease as well as SNP markers in linkage disequilibrium that correlate with specific responses to the medicine.

Pharmacogenetics and drug development

More efficient clinical trials and enhanced drug surveillance

Application of SNP mapping technologies will enable effective medicines to be developed and made available for clinical use more rapidly. Using abbreviated SNP linkage disequilibrium mapping, medicine response profiles could be identified during phase II clinical trials. These could be used in the selection of patient groups enriched for efficacy in phase III studies. This is likely to make these trials smaller, faster and more efficient⁴⁸.

Regulatory agencies would correctly be concerned that there were not enough patients in these streamlined phase III trials to evaluate adverse events, although larger clinical trials that do not select 'efficacy' patients are also unlikely to detect rare adverse events (less than 1 in 1,000). Regulatory authorities would also be apprehensive that, when the drug is marketed, patients who did not meet the pharmacogenetic criteria for prescription may be prescribed the drug without study of their potential benefits or adverse events. However, the risk–benefit ratio for patients with poor efficacy predictions may exclude them from phase III studies on ethical grounds as they would now knowingly be included solely to experience potential adverse events. Furthermore, in clinical practice, access to the medicine could be determined by prescriptions based on pharmacogenetic profiles.

In fact, pharmacogenetic technology may enable a significantly enhanced post-approval surveillance system to be established for approved medicines. Regulatory agencies, pharmaceutical companies and the public recognize the need to improve strategies for drug surveillance^{49,50}. In this proposed concept of regulated surveillance, hundreds of thousands of patients who receive the medicine would have blood spots taken and stored on filter papers in an approved location using the original blood sample screened for the initial medicine response profile for efficacy. As rare, serious adverse events are documented and characterized, DNA from patients who experienced the adverse event could be extracted and compared with DNA from control patients who received the drug but did not experience the adverse event. This would enable abbreviated SNP profiles for patients susceptible to the adverse event to be determined. These adverse event profiles would be combined with efficacy profiles to produce a comprehensive medicine response profile. This would allow selection of patients for both efficacy and lower complications of therapy (Fig. 4).

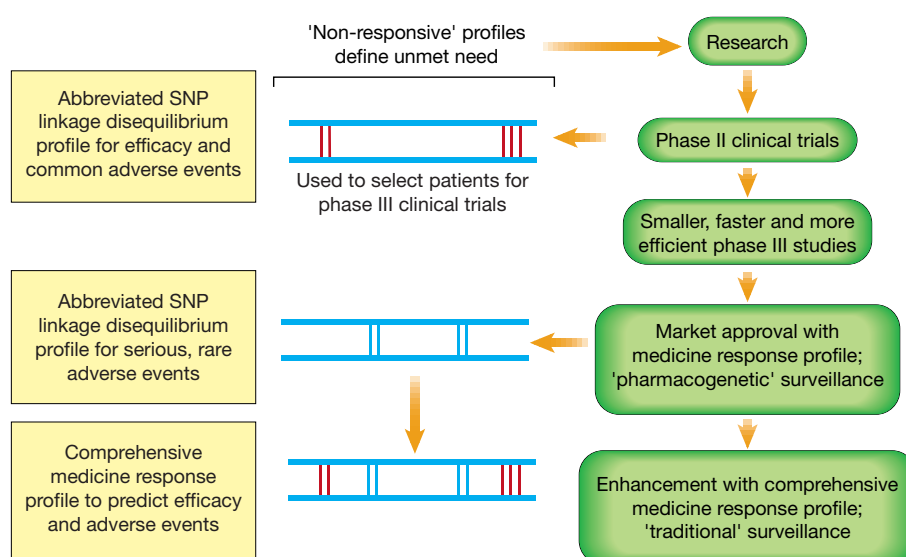
A predictive pharmacogenetic adverse event profile derived from hundreds of thousands of patients taking the drug would be a major advance on the present system of documenting reported serious adverse events during the use of the medicine in clinical practice, as this current system often obtains little or no predictive information to help subsequent patients, other than broad warnings.

Over the next few years, as we approach the ability to differentiate patients by their therapeutic responses, regulatory agencies and pharmaceutical companies will need to work together to pilot and examine methods to evaluate fewer total patients in faster, more efficient clinical trials while enhancing drug surveillance systems. Initial studies using medicine response profiles would no doubt use nested populations of patients within trials designed to meet current guidelines and regulations in order to demonstrate proof of concept.

Medicines for all

The application of pharmacogenetics will not diminish the population in whom a drug is effective, but simply allow prediction of patient response rather than prolonged and expensive prescribing by trial and error. Just as it will be possible to identify patients with drug efficacy, it will also be possible to identify those patients who do not respond early in the process of drug development. The ability to target heterogeneous groups of patients for parallel drug development early, rather than waiting years for non-responsive populations to emerge after

Figure 4 The development of a pharmacogenetic medicine response profile. An abbreviated SNP profile to predict efficacy could be identified in phase II clinical trials by detecting those SNPs along the genome that are in linkage disequilibrium when patients with efficacy are compared with patients who did not respond to the drug candidate. An abbreviated profile of these small regions of linkage disequilibrium that differentiate efficacy can then be used to select patients for larger phase III studies. This could make many of these phase III studies smaller and therefore more efficient. Pharmacogenetics could also be used during the initial post-marketing surveillance period to identify SNP markers associated with serious but rare adverse events. These markers could be added to the SNP markers for efficacy and common adverse events identified during development to produce a comprehensive medicine response profile, and to identify which patients respond to the drug and which patients will be at high risk for an adverse event.



extensive clinical use of the medicine, will be a significant benefit (Fig. 4). For example, SNP profiling of different medicine-responsive association groups during phase II trials will enable identification of the location of genes contributing to heterogeneous forms of the disease, leading to the discovery of new medicines and additional susceptibility targets.

By focusing clinical trials on patients who are most likely to respond, drug development resources could be targeted to those patients with continued unmet medical need. In particular, molecules that show less than a 30% response rate in a large population, but have clear efficacy in an identifiable smaller population of patients, would become viable as they could be readily identified for development and clinical practice.

As a result of disease heterogeneity, there may be large, definable sub-groups of patients suffering with a common phenotype, for example Alzheimer's disease, which represent only 10–15% of patients with that diagnosis. Focusing drug development on sub-groups of patients selected by either a disease-specific diagnostic or a medicine response profile will provide opportunities to develop more medicines for a larger proportion of patients with heterogeneous diseases. Similarly, patient groups who have vaguely defined phenotypes that are more difficult to categorize by objective criteria, such as depression, could be studied more efficiently using medicine response profiles as selection variables.

Value of pharmacogenetics to health-care delivery

The cost-effectiveness of new medicines (which are the product of considerable investment in research and development) is a significant concern to patients, funding bodies and governments^{46,51,52}. The application of pharmacogenetics to the delivery of medicines will maximize the value of each medicine. Medicines would be prescribed to only those patients where a high probability of efficacy without significant adverse events is expected^{45,46}. This is a much-preferred scenario than the problems facing funding agencies and governments at the present time. Medicines that might be prescribed to 100 patients to achieve an effect in 20 are becoming more difficult for sponsors of medical care to consider. However, selection of predicted responders offers a more efficient and economical solution to a growing problem that is leading governments and health-care providers to deny effective medicines to the few, because a

proportion of patients do not respond to the treatment⁵². The economy of predictable efficacy, limited adverse events, lower complications owing to targeted delivery, and increased cost-effectiveness of medicines will improve health-care delivery and eliminate the need for rationing. Effective and well-tolerated medicines with predictive medicine response profiles will obviate the need for formulary restrictions on prescribing and new policies to mandate cost-effectiveness to be proved in a broad population of patients.

Pharmacogenetics will impact medical care at multiple levels. As well-tolerated and effective medicines that treat, cure or prevent common diseases become a greater proportion of the medical care bill, the costs of chronic debilitating illnesses will be significantly reduced. As treatment and prevention of chronic and common diseases improves, a significant proportion of money saved by reducing hospitalization and long-term care costs could be transferred to well-tolerated and effective medicines.

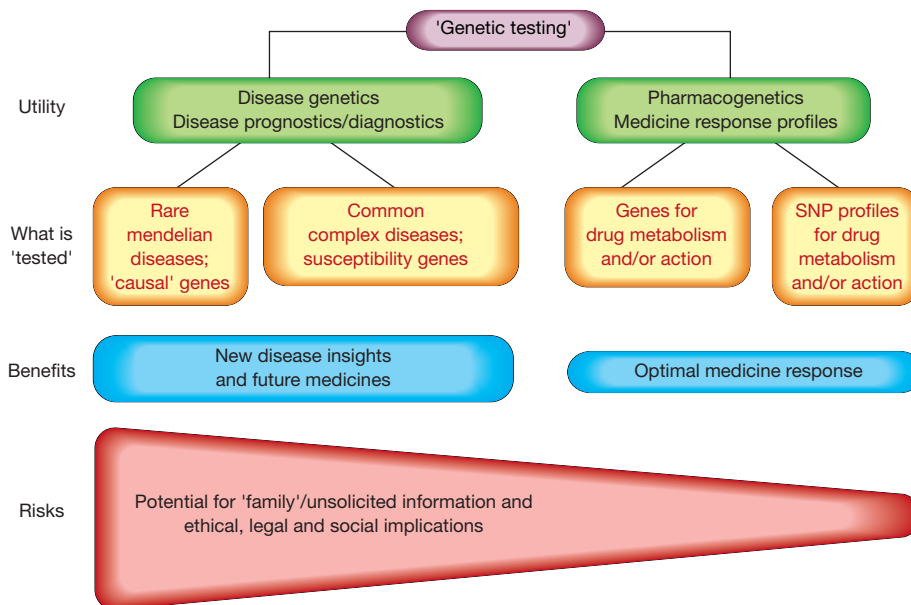
Understanding the differences in 'genetic testing'

The term 'genetic testing' is currently used indiscriminately to refer to very different applications of genetic science. It has entered into common vocabulary with very little specificity surrounding the wide diversity included in this shorthand term. Figure 5 illustrates some of the differences in using the term 'genetic testing'. Until now, government-sponsored committees convened to address 'genetic testing' have generally limited their definition and their reports to concerns regarding diseases caused by single-gene mutations. For example, the US National Institutes of Health Task Force on Genetic Testing and the SACGT (Secretary's Advisory Committee on Genetic Testing), its successor, have dealt mainly with mutational genetics and the need for government oversight in this area. While this objective has considerable merit, it represents only part of the spectrum of 'genetics tests'. Unfortunately, subsequent references to the Task Force conclusions, particularly by ethics commentators, have broadened the limited scope of the Task Force report⁵³. Quite distinct differences in recommendations for patients and relatives of patients with complex diseases are frequently miss-stated with authority by authors whose only experience is in mutational diseases^{53,54}.

Another class of 'genetic tests' is related to pharmacogenetics, including polymorphic detoxifying enzymes, drug-receptor variants or other inherited polymorphic traits that are not diagnostic of

Figure 5 'Genetic testing' needs to be defined carefully. The magnitude of the ethical, legal and social implications of genetic testing is dependent on the information derived from the test. Genetic tests for mutations in single genes that are causally related to rare diseases and are inherited in a simple mendelian fashion can have profound implications for the individual and family members. Genetic tests for disease-susceptibility gene polymorphisms — which are risk factors for the disease — have the added complication of uncertainty. In both cases the lack of effective intervention drives many of the issues. Pharmacogenetic profiles, on the other hand, will predict if an individual patient is likely to benefit from a medicine and be free of serious side effects. These profiles will not be designed to provide any other information, as the profile data are derived from the patients who respond with efficacy or adverse event when

taking the drug, compared with patients who did not respond. It does not differentiate disease. Should a polymorphism that is found to be related to disease association be included in a profile, it can be removed and replaced by another SNP that is in linkage disequilibrium, thus avoiding any disease-specific association, even if inadvertent. This would be similar to replacing the *ApoE4* SNP by one or more of the others in linkage disequilibrium with *ApoE4* but not specifically associated with Alzheimer's disease. The ethical, legal and social implications of pharmacogenetic profiles are therefore of a lower magnitude of societal concern compared with specific genetic tests for disease. (From ref. 46; published with permission.)



disease^{29–31,55}. In fact, when terms such as 'genetic testing' are applied, differentiation between tests that are specific to disease genes and profiles that are specific to genes involved in drug metabolism are often not well appreciated. Greater specificity of language is required to differentiate tests for disease genes from profiles for non-disease genes. Similarly, distinctions exist between non-disease gene polymorphisms associated with metabolic and drug-target characteristics and extended genomic profiles (for example, abbreviated SNP linkage disequilibrium profiles or medicine response profiles) that simply describe the phenotypic response (efficacy or adverse events) in response to a medicine.

Specificity of language use can be clarified in a hypothetical example. Assume that a 62-year-old man presents with symptoms of dementia that, after a thorough evaluation for other causes of dementia, is diagnosed as 'probable Alzheimer's disease'. If that patient carried an *APP717* mutation or an *ApoE4/4* homozygous genotype, the probability of accurate diagnosis of Alzheimer's disease, defined by subsequent autopsy neuropathologic confirmation, goes from 60–70% at clinical diagnosis to >97%^{26,57}. Both are disease-specific diagnostic 'genetic tests' and both provide predictive value in a symptomatic patient, although the *APP717* mutation is generally (but incorrectly) interpreted as being 100% predictive before symptoms begin. It should be noted that there are only two dozen families carrying autosomal dominant *APP* mutations associated with early-onset Alzheimer's disease. Most of these families segregate the *APP717* mutation. Thus there are less than 100 known *APP717* individuals carrying the *APP717* mutation. There are, however, three asymptomatic individuals who are at least one, or two, standard deviations over the mean age of onset for *APP717* mutations. All carry the *ApoE2/3* genotype. To date, no patient with the *APP717* mutation who developed clinical Alzheimer's disease has carried the *ApoE2/3* (or *ApoE2/2*) genotype. *ApoE2/3* seems to

protect from the *APP717* mutation. Thus, genetic counselling predictions made from measuring the *APP717* allele should not be made without also considering concomitant *ApoE* genotyping. These data and their significance have been either unknown to or perhaps unappreciated by 'ethics' commentators. *APP717* provides predictive information before any symptoms because it is very rare and disease begins in the 40–60-year age range^{17,58}. Carrying two *ApoE4* alleles does not predict Alzheimer's disease, only an increased susceptibility for the development of the disease as a function of age compared with other *ApoE* genotypes. Both are examples of disease gene-specific tests with very different implications for asymptomatic individuals, family members and societal risks of medical-care burden (Fig. 5).

Assume that a hypothetical drug exists for Alzheimer's disease that has two properties. The first is that the half-life of the drug in people varies as a function of a cytochrome P450 drug metabolizing polymorphism. The second property is a greater probability of efficacy in patients with a particular pharmacogenetic profile, that is, there is an abbreviated SNP profile using a panel of 400 SNPs from a map of 200,000 SNPs. 'Genetic testing' using the abbreviated SNP profile could select this particular drug for this patient, whereas a P450 'genetic test' might indicate a higher or more frequent dosing schedule. Neither provides any information about Alzheimer's disease. Neither provides any significant negative collateral information to relatives about Alzheimer's disease. Neither profile has the same ethical implications as measuring a mendelian mutation (*APP717*) or disease-specific susceptibility genotype (*ApoE4/4*). However, all of these are referred to as 'genetic tests'.

The abbreviated SNP linkage disequilibrium profiles will predict patients' responses to medicines, but they will not specifically 'test' the patient for the presence or absence of a disease gene-specific mutation, nor will they provide any other significant disease-specific predictive

information about the patient or family members. For practical purposes they would be anonymous laboratory profiles providing a read-out of predicted efficacy and adverse events. Medicine response profiles will simply measure phenotypic responses to a medicine based on a pattern of inherited factors detected as small regions of linkage disequilibrium. Thus, 'genetic' methods would be used to differentiate those patients who experience good efficacy and lower significant adverse events in response to a medicine from other patients who fail to respond or develop serious adverse events. The genetics of response to the medicine will be the only data generated using an abbreviated SNP linkage disequilibrium profile and, practically, could be easily designed, edited and safeguarded to be totally meaningless with respect to any known disease-specific gene information.

Traditional genetic counselling regarding education about disease inheritance would be of little value to an individual or a relative because no predictive information about disease risk is identified in the SNP linkage disequilibrium profile. Thus, as a practical matter, ethical and legal considerations of disease-specific gene tests, drug target or metabolic gene profiles, and abbreviated SNP linkage disequilibrium profiles for drug response deserve to be considered independently. As the scientific base shifts over the next decade from rare mutational to common diseases affecting millions of people, the rules governing 'genetic testing' should accurately reflect these distinctions. It is therefore incumbent that medical guidelines for mendelian- or susceptibility-gene testing do not extend automatically to discussions of other types of genetically based profiles in pharmacogenetics. Clear language and differentiation of respective ethical, legal and societal issues are required to prevent inaccurate vernacular usage creating a confused public perception of 'genetic testing'. □

1. Curran, M. E. Potassium ion channels and human disease: phenotypes to drug targets? *Curr. Opin. Biotechnol.* **9**, 565–572 (1998).
2. Marton, M. J. *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* **4**, 1293–1301 (1998).
3. Wiley, S. R. Genomics in the real world. *Curr. Pharmaceut. Des.* **4**, 417–422 (1998).
4. Blackstock, W. P. & Weir, M. P. Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* **17**, 121–127 (1999).
5. Kozian, D. H. & Kirschbaum, B. J. Comparative gene-expression analysis. *Trends Biotechnol.* **17**, 73–78 (1999).
6. Rockett, J. C., Esdaile, D. J. & Gibson, G. G. Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential. *Xenobiotica* **29**, 655–691 (1999).
7. Roses, A. D. Apolipoprotein E affects the rate of Alzheimer disease expression: beta-amyloid burden is a secondary consequence dependent on APOE genotype and duration of disease. *J. Neuropathol. Exp. Neurol.* **53**, 429–437 (1994).
8. Saunders, A. M. Apolipoprotein E and Alzheimer's disease: an update on genetic and functional analyses. *J. Neuropathol. Exp. Neurol.* (in the press).
9. Boschert, U., Merlo-Pich, E., Higgins, G., Roses, A. D. & Catsicas, S. Apolipoprotein E expression by neurons surviving excitotoxic stress. *Neurobiol. Dis.* **6**, 508–514 (1999).
10. Xu, P. T. *et al.* Human apolipoprotein E2, E3 and E4 isoform-specific transgenic mice: human-like pattern of neuronal immunoreactivity in central nervous system not observed in wild type mice. *Neurobiol. Dis.* **3**, 229–245 (1996).
11. Xu, P. T. *et al.* Regionally specific neuronal expression of human APOE gene in transgenic mice. *Neurosci. Lett.* **246**, 65–68 (1998).
12. Xu, P. T. *et al.* Specific regional transcription of apolipoprotein E in human brain neurons. *Am. J. Pathol.* **154**, 601–611 (1999).
13. Huang, J. T. *et al.* Interleukin-4-dependent production of PPAR- γ ligands in macrophages by 12/15-lipoxygenase. *Nature* **400**, 378–382 (1999).
14. Kliewer, S. A., Lehmann, J. M. & Wilson, T. M. Orphan nuclear receptors: shifting endocrinology into reverse. *Science* **284**, 757–760 (1999).
15. Barroso, I. *et al.* Dominant negative mutations in human PPAR γ associated with severe insulin resistance, diabetes mellitus and hypertension. *Nature* **402**, 880–883 (1999).
16. Roses, A. D. Apolipoprotein E alleles as risk factors in Alzheimer's disease. *Annu. Rev. Med.* **47**, 387–400 (1996).
17. Goate, A. *et al.* Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature* **349**, 704–706 (1991).
18. Levy-Lehah, E. *et al.* Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science* **269**, 973–977 (1995).
19. Sherrington, R. *et al.* Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature* **375**, 754–760 (1995).
20. Van Nostrand, W. E., Wagner, S. L., Haan, J., Bakker, E. & Roos, R. A. Alzheimer's disease and hereditary cerebral hemorrhage with amyloidosis-Dutch type share a decrease in cerebrospinal fluid levels of amyloid beta-protein precursor. *Ann. Neurol.* **32**, 215–218 (1992).
21. Aitman, T. J. & Todd, J. A. Molecular genetics of diabetes mellitus. *Baillieres Clin. Endocrinol. Metabol.* **9**, 631–656 (1995).
22. Sandford, A., Weir, T. & Pare, P. The genetics of asthma. *Am. J. Respir. Crit. Care Med.* **153**, 1749–1765 (1996).
23. Schork, N. J. Genetics of complex disease: approaches, problems, and solutions. *Am. J. Respir. Crit. Care Med.* **156**, S103–S109 (1997).
24. Concannon, P. *et al.* A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nature Genet.* **19**, 292–296 (1998).
25. Howard, T. D. *et al.* Genetics of allergy and bronchial hyperresponsiveness. *Clin. Exp. Allergy* **29**(Suppl.), 86–89 (1999).
26. Talley, N. J. Irritable bowel syndrome: disease definition and symptom description. *Eur. J. Surg.* **583**(Suppl.), 24–28 (1998).
27. Paterson, W. G. *et al.* Recommendations for the management of irritable bowel syndrome in family practice. IBS Consensus Conference Participants. *Can. Med. Assoc. J.* **161**, 154–160 (1999).
28. Hamm, L. R. *et al.* Additional investigations fail to alter the diagnosis of irritable bowel syndrome in subjects fulfilling the Rome criteria. *Am. J. Gastroenterol.* **94**, 1279–1282 (1999).
29. Evans W. E. & Relling, M. V. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **286**, 487–491 (1999).
30. Meyer, U. A. & Zanger, U. M. Molecular mechanisms of genetic polymorphisms of drug metabolism. *Annu. Rev. Pharmacol. Toxicol.* **37**, 269–296 (1997).
31. Cohen, L. J. & DeVane, C. L. Clinical Implications of antidepressant pharmacokinetics and pharmacogenetics. *Ann. Pharmacotherapy* **3**, 1471–1480 (1996).
32. Rudorfer, M. V., Lane, E. A., Chang, W. H., Zhang, M. & Potter, W. Z. Desipramine pharmacokinetics in Chinese and Caucasian volunteers. *Br. J. Clin. Pharmacol.* **17**, 433–440 (1984).
33. Mueller, R. F. & Young, I. D. in *Emery's Elements of Medical Genetics* (ed. Kalow, W.) 169–175 (Churchill Livingstone, Edinburgh, 1997).
34. Zhao, L. P., Aragaki, C., Hsu, L. & Quiaot, F. Mapping of complex traits by single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **63**, 225–240 (1998).
35. Brookes, A. J. The essence of SNPs. *Gene* **234**, 177–186 (1999).
36. Hacia, J. G. *et al.* Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nature Genet.* **22**, 164–167 (1999).
37. Marez, D. *et al.* Polymorphism of the cytochrome P450 CYP2D6 gene in a European population: characterization of 48 mutations and 53 alleles, their frequencies and evolution. *Pharmacogenetics* **7**, 193–202 (1997).
38. McKusick, V. A. Foreword to *Mendelian Inheritance in Man* 10th edn, vol. 1, xv–xxxvii (The Johns Hopkins University Press, Baltimore, 1992).
39. Lai, E., Riley, J., Purvis, I. & Roses, A. A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. *Genomics* **54**, 31–38 (1998).
40. Martin, E. R. *et al.* Analysis of association at single nucleotide polymorphisms in the APOE region. *Genomics* **63**, 7–12 (2000).
41. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
42. Collins, A., Lonjou, C. & Morton, N. E. Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl Acad. Sci. USA* **96**, 15173–15177 (1999).
43. Ott, J. Predicting the range of linkage disequilibrium. *Proc. Natl Acad. Sci. USA* **97**, 2–3 (2000).
44. Chen, J. *et al.* Microsphere-based assay for single-nucleotide polymorphism analysis using single base chain extension. *Genome Res.* **10**, 549–557 (2000).
45. Roses, A. D., Manasco, P. & Freeman, A. J. Pharmacogenetics and genetics to discover new medicines. *J. Comm. Biotech.* (in the press).
46. Roses, A. D. Pharmacogenetics and the future of drug development and delivery. *Lancet* **355**, 1358–1361 (2000).
47. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96**, 6745–6750 (1999).
48. Bonnie, A., Fijal, M. S., Hall, J. M. & Witte, J. S. Clinical trials in the genomic era. Effects of protective genotypes on sample size and duration of trial 2000. *Controlled Clin. Trials* **21**, 7–20 (2000).
49. Lazarou, J., Pomeranz, B. H. & Corey, P. N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *J. Am. Med. Assoc.* **279**, 1200–1205 (1998).
50. Wood, A. J., Stein, C. M. & Woosley, R. Making medicines safer—the need for an independent drug safety board. *N. Engl. J. Med.* **339**, 1851–1854 (1998).
51. White, J. Targets and systems of health care cost control. *J. Health Politics Policy Law* **24**, 653–696 (1999).
52. Mather, D. B., Sullivan, S. D., Augustein, D., Fullerton, P. & Atherton, D. Incorporating clinical outcomes and economic consequences into drug formulary decision: a practical approach. *Am. J. Man. Care* **5**, 277–285 (1999).
53. McConnell, L. M. B. *et al.* Genetic testing and Alzheimer disease: recommendations of the Stanford Program in Genomics, Ethics, and Society. *Genetic Testing* **3**, 3–12 (1999).
54. Koenig, B. A., Greely, H. T., McConnell, L. M., Silveberg, H. L. & Raffin, T. A. Genetic testing for BRCA1 and BRCA2: recommendations of the Stanford Program in Genomics, Ethics, and Society. Breast Cancer Working Group. *J. Womens Health* **7**, 531–545 (1998).
55. Taningher, M., Malacarne, D., Izzotti, A., Ugolini, D. & Parodi, S. Drug metabolism polymorphisms as modulators of cancer susceptibility. *Mutation Res.* **436**, 227–261 (1999).
56. Saunders, A. M. *et al.* Specificity, sensitivity, and predictive value of apolipoprotein-E genotyping for sporadic Alzheimer's disease. *Lancet* **348**, 90–93 (1996).
57. Mayeux, R. *et al.* Utility of the apolipoprotein E genotype in the diagnosis of Alzheimer's disease. Alzheimer's disease Centers Consortium on Apolipoprotein E and Alzheimer's disease. *N. Engl. J. Med.* **338**, 506–511 (1998). [Published erratum appears in *N. Engl. J. Med.* **338**, 1325 (1998).]
58. Roses, A. D. Genetic testing for Alzheimer disease. Practical and ethical issues. *Arch. Neurol.* **54**, 1226–1229 (1997).
59. Roses, A. D. Apolipoprotein E is a relevant susceptibility gene that affects the rate of expression of Alzheimer's disease. *Neurobiol. Aging* **15**(Suppl 2), S165–S167 (1994).

Acknowledgements

I acknowledge my colleagues at Glaxo Wellcome in the development of the strategies outlined in this review; in particular, the vision and support of J. Nield and the commitment of the Glaxo Wellcome R&D Executive Committee — A. Baxter, T. Eaves, A. Hennah and J. Palmer — in using their wealth of experience to help make this vision become a reality. I also thank the Genetics Directorate at Glaxo Wellcome for their expertise and continued hard work. Some of the recent pre-publication work carried out by Glaxo Wellcome scientists is reviewed in this paper. I especially thank A. Freeman, A. Saunders and A. Kidgell for expert editorial assistance.

Aventis and functional genomics



In mid-2000, at the beginning of a new century, no topic could be more appropriate for Nature Insight than functional genomics. Of all the scientific disciplines, none will have a greater influence on the future of drug discovery than genomics, and none will change medicine more profoundly.

Exactly one hundred years have passed since the genetic rules of Gregor Mendel, originally published in 1865, were rediscovered in the reports of the German Botanic Society, their significance having initially been overlooked. In 1900

Mendel's findings flourished in a fertile environment of scientific curiosity, which laid the foundations for the subsequent genetic revolution of the 20th century.

At the height of this century, the world was struck by Watson's and Crick's discovery of the DNA structure and its ramifications: "It has not escaped our notice that the specific pairings we have postulated immediately suggest a possible copying mechanism for the genetic material", they concluded in an article published in "Nature" announcing the news to the world. Along with the digital revolution, the genetic revolution has already radically changed the face of science and propelled mankind forward towards its roots, while biology is being transformed more and more into an information science.

Sequencing the human genetic information is a task that has now almost been completed. The speed at which this has been achieved goes beyond all expectations of just 12 years ago. It has taken us into the entrance hall of the library of life, whose still largely incomprehensible volumes contain the three billion letters of our genetic make-up, now waiting to be joined together into meaningful words, sentences, paragraphs and chapters. With the resulting instructions we shall be

able to make life on earth more worthwhile.

Such a vast decoding task makes deciphering the hieroglyphics look like child's play. To accomplish this task, we all need to combine the urge for scientific knowledge with an unprecedented responsibility for life and its preservation. This library of life affects the subjective life of us all to a greater extent than any scientific factors have ever done in the past. If we use it wisely, we shall have a unique opportunity to ensure the sustainable development of life on Earth.

"Our challenge is Life" is the motto that governs our actions here at Aventis. Functional genomics – together with a wide array of enabling technologies – plays a central part in meeting this challenge, and we, as a leading life sciences company, are supporting it with all available resources. We hope the results achieved in the near future will help millions of patients throughout the world by healing conditions for which there has so far been no adequate treatment, if any at all.

A handwritten signature in blue ink that reads "Frank L. Douglas M.D.".

Frank L. Douglas
Executive Vice President
Head of Drug Innovation & Approval

Electrophoresis for gels

Windows but no pain, doors but no entry, Watson but with Crick.

Modular electrophoresis

Flowgen <http://www.flowgen.co.uk>

Modular flexibility in electrophoresis

With Flowgen's latest modular vertical gel electrophoresis system, one basic unit can perform slab gel electrophoresis (with hand-cast or pre-cast gels), capillary gel electrophoresis and western blotting. The system is available in both 10 × 10 cm and 20 × 20 cm formats. Two slab gels, ten capillary gels or two gels for blotting can be run in the same outer running tank. The electro-blotting module performs western and other types of blotting. The electrodes are only 6 cm apart for maximum current strength.

Reader Service No. 100

New Electrophoresis Range

Hybaid Limited <http://www.hybaid.com>

High-throughput electrophoresis products

In February this year Hybaid introduced a new range of electrophoresis gel tanks and accessories designed for the high-throughput laboratory. The use of microtitre combs allows up to 400 samples to be loaded using a multi-channel pipette. Features include easy-loading, UV transparent trays with fluorescent ruler, external casting and combs to suit most purposes. Hybaid is also offering a wider range of high purity agarose gels, molecular weight markers and loading buffers.

Reader Service No. 101

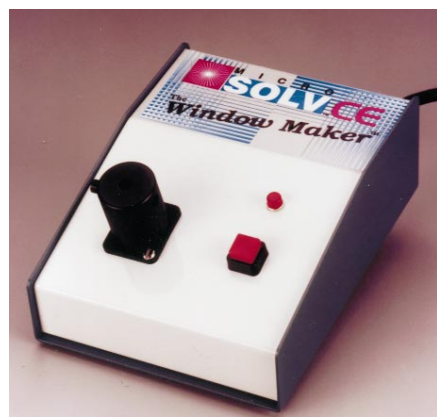
Window Maker

MicroSolv Technology Corporation

<http://www.microsolvtch.com>

Look through any window

Designed for the controlled stripping of polyimide to create detection windows in protectively clad, fused silica capillaries, the Window



Watch it: MicroSolve put you close to the action.



Vertical line: Flowgen's tanks on parade.

Maker creates a reproducible 2-mm window area on both coated and uncoated capillaries (it is not recommended for use with polyacrylamide or CE-100SA, sulphonic acid coated capillaries). The idea behind a 2-mm window is that the column retains its strength while providing an adequate area for detector line-up. The Window Maker is available for 110 volt or 220 volt supplies and can be supplied with country specific plugs as required.

Reader Service No. 102

Gel Pro

Media Cybernetics <http://www.mediacy.com>

For the image conscious

The Gel-Pro Imager Kit is a multipurpose system for acquiring, archiving, documenting, analysing and reporting digital images from electrophoresis experiments. Applications include imaging of DNA, protein and membrane samples involved with electrophoresis, with a claimed saving of time over other methods. The kit features a camera system with a LCD monitor and controller built into the enclosure unit. The entire unit weighs less than 20 pounds and mounts on any transilluminator. The kits include Gel-Pro Analyzer software which enables qualitative molecular weight analysis, quantitative amount analysis and reporting for gels and dot/slot blots.

Reader Service No. 103

NucleoScan 2000

NucleoTech <http://www.nucleotech.com>

Fluorescent primers, who needs 'em?

This real-time gel electrophoresis and imaging system has been developed to produce reproducible high resolution images of electrophoretic gels. An on-board solid-state laser (532 nm) allows excitation and detection of a

wide spectrum of fluorescent dyes including ethidium bromide, avoiding the need for expensive labelled fluorescent primers. The NucleoScan 2000 can be used for genotyping, microsatellite analysis, AFLP, SSCP, mutation screening, quantitative PCR and sequencing.

Reader Service No. 104

VectaSpin 3

Whatman <http://www.whatman.com>

Two-stage filtration

A pigment-free polypropylene centrifuge tube filter with a 3-ml capacity, VectaSpin 3 is a combination filter designed for protein purification. The top layer acts as a pre-filter, removing coarse particulates and preventing clogging, whilst the lower layer is a fine filter (0.7 µm) to remove finer particulate matter and act as an oleophobic barrier to retain solvent in the well until vacuum is applied. This double-filter technology is used in other Whatman products, including a system designed for high-throughput clean-up in HTS autoinjectors.

Reader Service No. 105

ADVERTISEMENTS

Custom Peptides & Polyclonal Antibodies

Please contact us about complete services for your custom peptide and polyclonal antibody needs.

Research Genetics

U.S. or Canada 800-533-4363

FAX 256-536-9016

Check out our homepage at
<http://www.resgen.com>

READER ENQUIRY NO. 39

Silver Stain Kit II for Electrophoresis

Rapid, sensitive detection of proteins & nucleic acids

Wako

Wako BioProducts
Wako Chemicals USA, Inc.
1600 Bellwood Road
Richmond, VA 23237
Telephone: (804) 271-7677
Facsimile: (804) 271-7791
(800) 992-9256
bioproducts@wakousa.com

READER ENQUIRY NO. 41

new on the market

pHlash

Genomic Solutions <http://genomicsolutions.com>

Gels in a pHlash

The Investigator/pHaser 2-D Gel Electrophoresis System introduces a new generation of pHlash IPG strips for isoelectric focusing. In this immobilized pH gradient chemistry, developed as an alternative to isoelectric focusing (IEF) with carrier ampholytes, the ampholytes are covalently linked to the acrylamide and the gradient is fixed in the gel. Advantages claimed for immobilized pH gradient strips include ease of use, less sensitivity to protein load and contaminants such as salts, and good resolution for soluble proteins. pHlash strips are available in four overlapping narrow pH ranges (pH 3–5, 4–8, 5–7 and 7–10) and one broad range (pH 3–10).

Reader Service No. 106

Electro-Fast Marker

Abgene <http://www.abgene.com>

Make a mark

Electro-Fast gel systems offer 96-sample throughput, with additional wells for markers, in a format suitable for multi-channel loading. They are intended for high throughput with samples that contain a small number of signals, such as PCR reactions. At least 96 samples can run on a single gel. Since the Electro-Fast has short separation distances, conventional marker bands can overlap making product size and quantitation calculations difficult. Hence the Electro-Fast Marker, consisting of three DNA bands of 200, 500 and 1,000 bp. The marker contains a red dye and precipitant for quick and easy loading on agarose gels.

Reader Service No. 107

Lipoprotein electrophoresis

C.B.S. Scientific <http://www.cbsscientific.com>

A touch of class

The new C.B.S. electrophoresis system is designed for the resolution of lipoprotein subclasses (LDL, HDL) based on particle size, using non-denatured pore gradient gel electrophoresis (GGE). Capable of accommodating up to four separate gels, the PGGE Pore Gradient Lipoprotein System is also suitable for standard PAGE and short DNA sequencing runs. The PGGE system can run between 1 and 4 pre-cast or manual cast gels with dimensions of 82 × 82 × 5 mm.

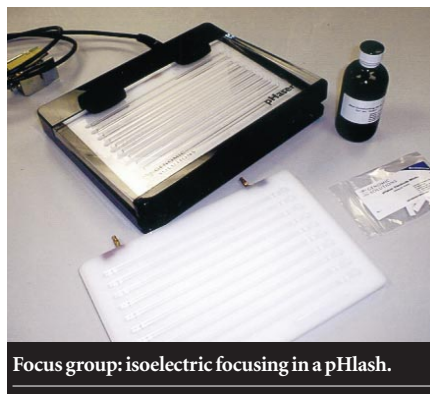
Reader Service No. 108

AlphaEase 5.5

Alpha Innotech <http://www.alphainnotech.com>

Upgrades free to existing customer's from the company's web site

The latest release of AlphaEase image analysis software, version 5.5, now includes the AutoLane function which automatically detects lanes and bands in a gel image with a



Focus group: isoelectric focusing in a pHlash.

single button click. The AutoSpot function allows for automatic two-dimensional detection and quantification of spots on a 2D gel image, and ChemiDisplay performs automatic contrast display adjustment for low light chemiluminescence imaging applications. Image Stacking in the MovieMode function allows for real time image addition and has the ability to stack one image frame to another as the ChemiImager or FluorChem camera acquires images.

Reader Service No. 109

Savant Primo

ThermoQuest SEG <http://www.tmquest.com>

Are Watson and Crick still in the picture?

Savant's new range of submarine gel chambers is made from a new plastic called Primo. The Primo tray is heat resistant, allowing the user to pour agarose at 90 °C without tray warping or crazing. The tray is resistant to many alcohols and acetone, making it suitable for RNA analysis. The Savant 250B power supply is recommended for the Primo range. Early purchasers of the combination of 250B and Primo chamber were offered a free poster of the painting 'Homage to Watson and Crick' — supplies were no doubt limited so check with ThermoQuest to see if it's still available.

Reader Service No. 110

EZ-Reach Door

Ultra-Lum <http://www.ultralum.com>

Cover story

This purpose-built door will fit 8 × 10 inch camera hoods to allow easy access to electrophoretic gels up to 20 × 20 cm. EZ-Reach eliminates awkward lifting of bulky hoods, and reduces the possibility of damage to fragile cameras. The door magnetically attaches to the hood maintaining a light-tight seal for photography. Any fluorescent or visibly stained gel up to 20 × 20 cm fits through the opening in the hood. Any of Ultra-Lum's gel documentation cameras can be mounted on the universal camera mounting bracket available for the EZ-Reach hood.

Reader Service No. 111

These notes are compiled in the Nature office from information provided by the manufacturers.